REPORT DOCUMENTATION PAGE

1. Recipient's Reference	2. Originator's Reference	3. Further Reference	4. Security Classification of Document
	AGARD-LS-130	ISBN 92-835-1459-9	UNCLASSIFIED

5. Originator: Advisory Group for Aerospace Research and Development
North Atlantic Treaty Organization
7 rue Ancelle, 92200 Neuilly sur Seine, France

6. Title: DEVELOPMENT AND USE OF NUMERICAL AND FACTUAL DATA BASES

7. Presented at: Gaithersburg, Maryland, USA on 5–6 October 1983
in London, UK on 10–11 October, 1983 and
in Lisbon, Portugal on 13–14 October, 1983.

8. Author(s)/Editor(s)	9. Date
Various	October 1983

10. Author's/Editor's Address	11. Pages
Various	130

12. Distribution Statement: This document is distributed in accordance with AGARD policies and regulations, which are outlined on the Outside Back Covers of all AGARD publications.

13. Keywords/Descriptors

Data bases
Information systems
Data processing
Numerical analysis
Information retrieval
Information theory
Systems engineering

14. Abstract

Lecture Series No.130 is concerned with the development and use of numerical and factual data bases, and is sponsored by the Technical Information Panel of AGARD and implemented by the Consultant and Exchange Programme.

Numerical and factual data, as sources of information for all levels of aerospace and defence R & D management and staff activity, are becoming increasingly important. These data are necessary to support research and engineering efforts in all fields. They are also becoming increasingly important to support of assist in the decision-making process. Today, a number of numerical data bases are available through national information centres and others are available from academic or commercial information sources. Data in many of these data bases can be retrieved and manipulated in display systems currently available. There is, however, a great need to improve the quality, reliability, availability, accessibility, dissemination, utilization and management of these data.

Better knowledge regarding the generation and availability of such data bases, and the techniques for their use, will be of benefit to the R & D community and their information service centres.

The scope of the Lecture Series includes: generation of numerical data, consideration of the quality and reliability of the data, methods for publishing and disseminating the data, a review of the data bases that are currently available, how these data bases can be used, and future needs for numerical data bases.

AGARD-LS-130

AGARD-LS-130

AGARD

ADVISORY GROUP FOR AEROSPACE RESEARCH & DEVELOPMENT

7 RUE ANCELLE 92200 NEUILLY SUR SEINE FRANCE

AGARD LECTURE SERIES No.130

Development and Use of Numerical and Factual Data Bases

NORTH ATLANTIC TREATY ORGANIZATION

DISTRIBUTION AND AVAILABILITY
ON BACK COVER

NORTH ATLANTIC TREATY ORGANIZATION

ADVISORY GROUP FOR AEROSPACE RESEARCH AND DEVELOPMENT

(ORGANISATION DU TRAITE DE L'ATLANTIQUE NORD)

AGARD Lecture Series No.130

DEVELOPMENT AND USE OF NUMERICAL AND FACTUAL DATA BASES

The material in this book has been assembled to support a Lecture Series under the sponsorship of the Technical Information Panel and the Consultant and Exchange Programme of AGARD, presented on 5–6 October 1983 in Gaithersburg, Maryland, USA; on 10–11 October 1983 in London, UK and on 13–14 October 1983 in Lisbon, Portugal.

THE MISSION OF AGARD

The mission of AGARD is to bring together the leading personalities of the NATO nations in the fields of science and technology relating to aerospace for the following purposes:

- Exchanging of scientific and technical information;
- Continuously stimulating advances in the aerospace sciences relevant to strengthening the common defence posture;
- Improving the co-operation among member nations in aerospace research and development;
- Providing scientific and technical advice and assistance to the North Atlantic Military Committee in the field of aerospace research and development;
- Rendering scientific and technical assistance, as requested, to other NATO bodies and to member nations in connection with research and development problems in the aerospace field;
- Providing assistance to member nations for the purpose of increasing their scientific and technical potential;
- Recommending effective ways for the member nations to use their research and development capabilities for the common benefit of the NATO community.

The highest authority within AGARD is the National Delegates Board consisting of officially appointed senior representatives from each member nation. The mission of AGARD is carried out through the Panels which are composed of experts appointed by the National Delegates, the Consultant and Exchange Programme and the Aerospace Applications Studies Programme. The results of AGARD work are reported to the member nations and the NATO Authorities through the AGARD series of publications of which this is one.

Participation in AGARD activities is by invitation only and is normally limited to citizens of the NATO nations.

Published October 1983

ISBN 92-835-1459-9

Printed by Specialised Printing Services Limited
40 Chigwell Lane, Loughton, Essex IG10 3TZ

PREFACE

This Lecture Series, concerned with the development and use of numerical and factual data bases, is sponsored by the Technical Information Panel of AGARD and implemented by the Consultant and Exchange Programme.

Numerical and factual data, as sources of information for all levels of aerospace and defence R & D management and staff activity, are becoming increasingly important. These data are necessary to support research and engineering efforts in all fields. They are also becoming increasingly important to support or assist in the decision-making process. Today, a number of numerical data bases are available through national information centres and others are available from academic or commercial information sources. Data in many of these data bases can be retrieved and manipulated in display systems currently available. There is, however, a great need to improve the quality, reliability, availability, accessibility, dissemination, utilization and management of these data.

Better knowledge regarding the generation and availability of such data bases, and the techniques for their use, will be of benefit to the R & D community and their information service centres.

The scope of the Lecture Series includes: generation of numerical data, consideration of the quality and reliability of the data, methods for publishing and disseminating the data, a review of the data bases that are currently available, how these data bases can be used, and future needs for numerical data bases.

LIST OF SPEAKERS

Lecture Series Director: R.R.Taschek
2035 47th Street
Los Alamos, NM 87544
USA

SPEAKERS

A.J.Barrett
ESDU International Ltd
251/9 Regent Street
London W1R 7AD
England

D.Lide, Jr
Office of Standard Reference Data
National Bureau of Standards
Washington, D.C. 20234
USA

D.G.Watson
Crystallographic Data Centre
Lensfield Road
Cambridge CB2 1EW
England

M.Chase
Dow Chemical Company
Thermal Research Laboratory
Midland, Michigan 48640
USA

J.Sutton
Scientific and Technical Information Unit
Department of Industry
Ebury Bridge House
2–18 Ebury Bridge Road
London SW1W 8QD
England

J.H.Westbrook
Materials Information Services
General Electric Company
Corporate Research and Development
120 Erie Boulevard
Schenectady, NY 12305
USA

CONTENTS

STATUS AND FUTURE OF THE USE OF NUMERICAL AND FACTUAL DATA

Dr. Richard F. Taschek
Los Alamos National Laboratory (Retired)
Mailing Address: 2035 47th Street
Los Alamos, NM 87544 U.S.A.

SUMMARY

The nature of numerical and factual data and data bases is discussed and their place in modern industrial society is reviewed. The components of the process by which data is produced, made public, extracted, assessed, evaluated and finally put to use, either as a single item or part of a data base, are examined. The present status of this process (the data cycle) is described and shortcomings briefly analyzed. The costs and organizational structures of various components are crudely evaluated and it is decided that the data generation component dominates cost, while the evaluation component is least well developed. A detailed data cycle flow chart is presented for guidance in future more detailed approaches to critical data generation and utilization.

1. Introduction

It is the purpose of the lecture series that we are about to begin, to discuss the whole of what will be called the data cycle. This will include all or most of the functions which are involved in the "development" of "data bases" in a mainly generic manner. When we have established generally how to develop data bases to order, the user interests will be discussed in detail while we already now agree that the user is the ultimate requestor-recipient of the data base, putting it into the context for which the whole exercise was carried out.

Let us now inquire into the natural evolution of this activity to acquire various insights. In the not too distant past the requirement for data arose when there was a problem needing solution. The datum or data required might be numeric (with units of course) for a scientific, technical or engineering problem or it might be informational/factual non-numeric for a problem which might arise in a demographic, economic, legal or other situation; or there might be a combination of numeric and factual data necessary to address the problem at hand. We will be dealing primarily, but not entirely, with numeric data.

In that earlier time mentioned above, the data user (or needer) would get most or all that he required, from the scientific literature, tables, or handbooks already in existence, but as the sophistication of his needs evolved two immediate difficulties arose, first that the data needed did not yet exist or second, that even if it did exist it did not satisfy the user's needs for accuracy, or range or morphology or cost or something else perhaps, usually situations dealing with independent variables. In many cases the user had the added difficulty that if new data had to be generated or old data improved, it was entirely out of his personal experience. Furthermore, the rapid advances of technology and engineering, initially based on comparatively small requirements for data, produced great changes in data applications as the Edisonian era disappeared. The design, fabrication, testing, and use of the rapidly evolving products of science, technology, engineering and manufacture became complex and sophisticated first with data sets from a single discipline needed, but soon from multidisciplinary sources. Then sharpened criteria for data quality became necessary and even urgent, particularly in sensitive design situations. Then the magnitude of the data set or sets needed to solve some of the larger problems expanded greatly and "data bases" pertinent to specific problems were born.

2. Definition and Use of Data Bases and Data Cycle

We may describe a data base as the useable form of scientific/technical information from one or more disciplinary subdisciplines, usually focused on a single area of application. The content may be in numeric or information form or both. The data base may simply exist in a notebook or it may be designed for computer usage and contain the necessary algorithms for its specific utilization, as a part of the data base, but not data in themselves. For many cases two or more computer utilizable data bases may be coupled for solving a larger or more complex problem.

The nature of the problem solving may be such that the data base needs only to be vertical (or one dimensional) e.g. the properties of materials - or it may be multidimensional as for instance in nuclear reactor design of core, design of shield, design of coolant system requiring nuclear data of several kinds, materials data, radiation damage data, chemical thermodynamic data, etc. The first situation may be coverable by a single data base, the latter by several interlocking bases. Fortunately the development of automated data processing was and is reasonably coordinated with the growing requirements for data handling.

It is very important, however, to remember that ADP is a tool for the manipulation of data and does not contribute to the substantive content of the data. Its great manipulative power is, however, a major factor in making possible a necessary

systematic approach to the totality of the data field and its utilization. This process is just beginning to fulfill its capabilities and forecasts major impacts in many areas. In particular, the coupling of automated data bases with automated and robotic production and manufacturing, just getting off the ground will place great emphasis on the quality of the data and data bases.

Data and data bases may originate from a variety of sources and are utilized by an even larger number of user groups. It is thus important to understand why and how data is produced and what the costs of the various components of the data cycle are since cost benefit analysis formal or not will finally determine the part to be played by various data bases in the application being considered.

3. The Place of Academia, Private Enterprise, and the Government in the Data Cycle.

The field of scientific, technical and engineering data is a highly fragmented activity, well served in some aspects, poorly done or not at all in others and often characterized by the lack of coherence in the overall system. The lack of a coordinated approach to this multifaceted subject is due partly to a lack of recognition of all the components by the three principal participating groups in the data cycle i.e. academia, the industrial/commercial sector and the public sector as represented mainly by the federal government in the US.

Modern industrial nations manufacture products and devices for the market place and are highly sensitive to the response of the market place. To be competitive with each other and internally within a nation, they must more and more be on the engineering forefront of rapidly developing technological situations. Motivations driven by the market place cause the competitive economics of production to be a major force in new developments even when the scientific and technical problems of an initiative are fully known and solved. Thus quality assured data and data bases must exist, be accessible and in useful formats as sine qua non factors, but economics of the free enterprise system may also determine that the initiative cannot now be pursued. Note that the motivations of the public sector may come to quite a different conclusion, e.g. in the obvious case of national defense when cost becomes a secondary issue to the solution of the problem.

In the private sector new or improved data can be produced through a particular industrial group's own efforts either in-house or by external contract. Since WWII this approach is followed primarily by larger industries capable of mounting research and development efforts having continuity and allowing comparatively far-off maturation time. Systematic continuing applied research and development efforts are rather rare in industry because of the high capital and effort costs of data production or generation, with no guarantee that it will be immediately useful in problem solving, design or production efforts. How then is this obviously necessary function of data generation achieved? The answer is, in part, that the public foots the bill for much of data generation as a national overhead cost necessary for the well being of the nation as a whole. Obviously this can only be a public expenditure and responsibility when done for the national defense function, which in turn however, generates very considerable scientific and technological data most of which becomes available to the user community as a whole through the publication in open professional journals or by means of in-house documents available from government printing facilities.

As implied above, the profit motive is not a primary factor in a large share of data generation and the remainder of the data cycle functions performed for the federal government. Aside from difference in detailed implementation, the generation of scientific and technical data in the US has since the end of WWII been taken over in a major way, approximately 90%, by departments and agencies of the federal government, especially by the Department of Energy and its predecessors, the Department of Defense, the Department of Agriculture, the National Institutes of Health for the Bio-sciences, the National Aeronautics and Space Agency, and for primarily academic research support, the National Science Foundation. This listing is not at all complete of course, viz the National Bureau of Standards in the Department of Commerce and the US Geological Survey in the Department of Interior which organizations considerably predate WWII.

Note that the data generated for and by these organizations is in response to their own mission-oriented assignments or is mandated by the Congress in specific statutes; thus the new data and information produced as objectives of the department becomes simultaneously a matter of public record for use by anyone. It is particularly important to remember that the numerical and informational data produced by these publically funded organizations is the quantitative description of what has been accomplished by their scientific and technical endeavors.

It will develop, if it is not already obvious, that by far the most costly function of the datacycle is the production or generation of the data in the physical sciences where it is fairly readily traceable; a similar situation exists in the biological research sciences and it is surmised that the same is also true in the factual data production and gathering of the socio-technical, socio-economic, demographic and similar data programs. To reiterate then, a great part of the numerical and factual data that gets incorporated into various data bases originates from a multiplicity of sources initially intended for narrowly focused objectives and nearly all supported by public funds. This most costly component of the data cycle is generally accepted as an

overhead charge to the national treasury, but the economic benefit is often not recognized nor is it easily traceable because the user community is so widely dispersed and because appreciable portions of the data generated may not be used in the near term (10 years) or perhaps never. On the other hand when specific data is urgently needed, it would be desirable to "have it on the shelf" especially if it were of a kind requiring one to ten years for generation. This would require a decidely different philosophy and approach to what is now in effect.

The data generating capabilities integrated over all of the government owned-government operated laboratories in the nation, plus the government funded non-mission research in academia or research institutes are very powerful indeed. It is becoming urgent to examine means by which the private sector might draw on these public facility capabilities for their own purposes, but without jeopardizing the requirements already assigned to these facilities. Data generation for use a decade or so in the future can likely be left to the present system, but quick response, and quick turn around must be more highly systematized, especially for multiple user interests.

Although it seems unlikely that major versatile data generating facilities will become common in the private sector, economic analyses of the remaining components of the data cycle may well indicate that many of these could be accomplished adequately within the market place.

4. Long Range Planning in the Technical Data and Information Fields

As user pressures for better and better data with more rapid recycling times and for the competent assemblage of data bases responsive to the specialized needs of each user group becomes apparent it will also become clear that a management structure compatible with the requirements will have to be put in place.

For purposes of assessing several matters, e.g. cost-benefit of evaluated data, data base assemblage, data management organizations, data production problems etc. it is necessary to examine the data cycle as it now exists, having grown mainly like Topsy. The conventional set of functions for the data sequence is quite evident and consists of the following for a programmatic objective. (Fig. 1)

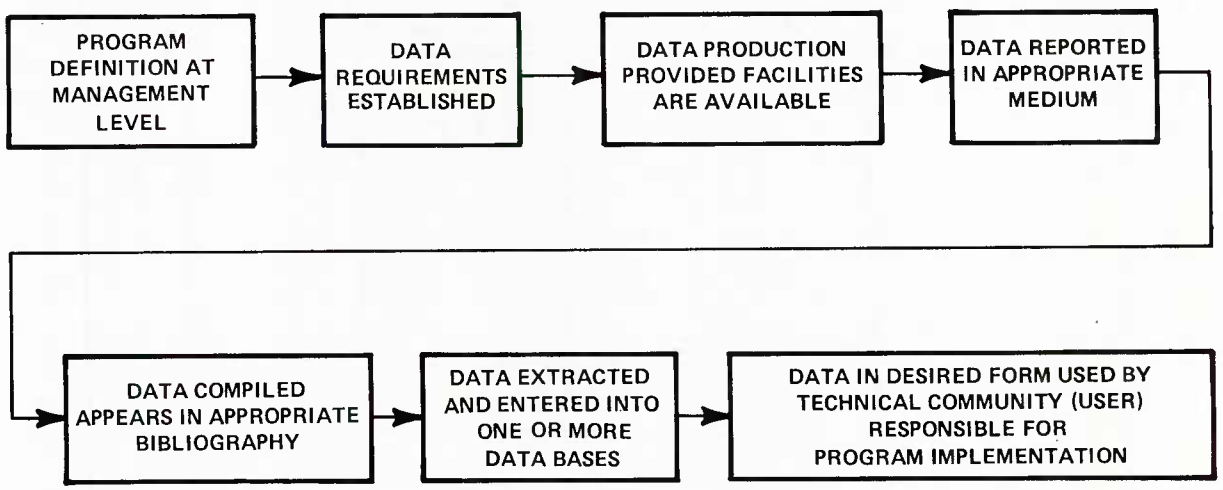


Fig. 1

Much of the time this sequence can be abbreviated to a bare bones form. Namely (Fig. 2).

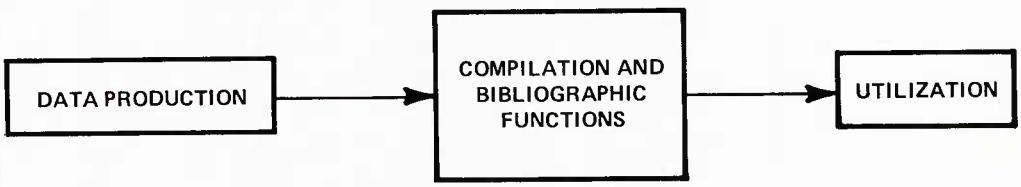


Fig. 2

With some exceptions, the execution of the steps listed in Fig. 1 occur in a quite laissez faire way rather than in the deliberate organized approach which would be used if there were a real emergency dependence on the data (or perhaps a major industrial advantage to have the data in hand).

When the data sequence is stated as in Fig. 2 it is quite evident that the determination of data quality (critical evaluation) is a missing factor of great importance to modern competitive technology and engineering and may even be a matter of life and death for sensitive military systems, or toxic waste disposal. If evaluation of the data is included, the sequence becomes (Fig. 3).

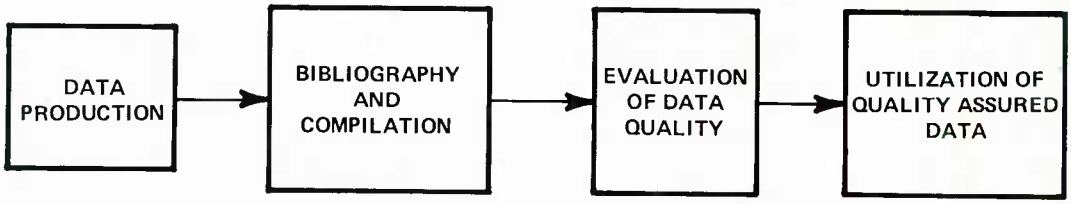


Fig. 3

It is unfortunate that the three step sequence of Fig. 2 is most commonly used, for a variety of reasons, the most cogent one probably being that critical evaluation requires fully professionally qualified scientists knowledgeable in detail about the way the data is generated and the associated difficulties. It is not unusual that the data producing scientists themselves become professional evaluators.

When the evaluation step is included as in Fig. 3, it now becomes possible for the sequence to be converted into a cycle. This arises from the fact that the evaluation allows a determination to be made as to whether the data is of adequate quality for the one or more programmatic implementer's purposes. A skeleton version of the data cycle now becomes (Fig. 4).

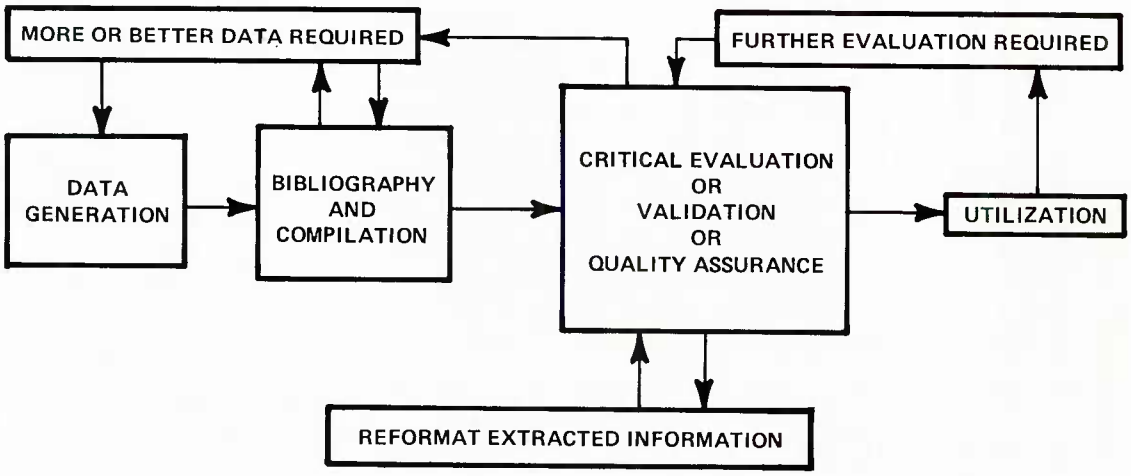


Fig. 4

5. Implementation of a Rudimentary Data Cycle

To accomplish this kind of activity efficient organization becomes desirable and necessary to optimize the interactive steps amongst functions since there must now be knowledgeable communication between non-adjacent activities. The data production and handling sequences and cycles so far presented are basically generic in nature, but at least the generators and users (perhaps also the evaluators) are almost invariably involved in a specific data area, nuclear physics, chemical thermodynamics etc. Each of the specialist groups may use the generic cycles, either deliberately or subconsciously, as guidance to their own modus operandi albeit the guidance may be rather close. While generators, evaluators and users each form a matrix column of functionally related units, the compilational and bibliographic requirements of many such sharply focused groups can be served by a single "collative information center." Realistically, a system of this kind is not so clearly defined, for there will be inevitably redundancies at the generator level and the multiple users of the same data unit output may not all be related. Similar comments may be true for the construction of broadly based data bases, especially those which are multidisciplinary and this will no doubt be discussed with more care by later lecturers.

To formulate and implement a model for the generic data cycle that is broadly

adequate to the complexities of highly sophisticated technologies, useful to multidisciplinary programs being carried out by non-experts in the data field and that is capable of establishing a degree of validity which forces acceptance also in legal and economic determinations, requires that issues and criteria be carefully set. This latter can only be done with the requisite degree of consistency by a properly empowered overview group, which will not only perform the day to day management of the particular specialized work unit, but monitor its accomplishments and provide quality assurance of the work done. The groups must have representatives of each of the primary functions in the model as it is applied to the particular data unit assigned. To illustrate what is meant by these comments a small segment of a much larger total data area is shown (Fig. 5) with likely interconnections. One category might be labeled Nuclear Reactor Technology. A similar segment of a large area could be set up for say, Geothermal Power Technology. The point to note is that each of the data generating units is composed of one or more specialists interacting only peripherilly with any of the others, but the data producer areas were so selected that evaluation of at least two groups could be done by a single person or a few. Furthermore, the utilization categories were so chosen as to illustrate that their input data came from several of the data-producer units presumably through the appropriate evaluators. In Fig. 5.1 a mirror image is shown for the category of Geothermal Power Technology with similar cross linkages. Of considerable interest are the additional direct linkages with the Nuclear Reactor Technology category.

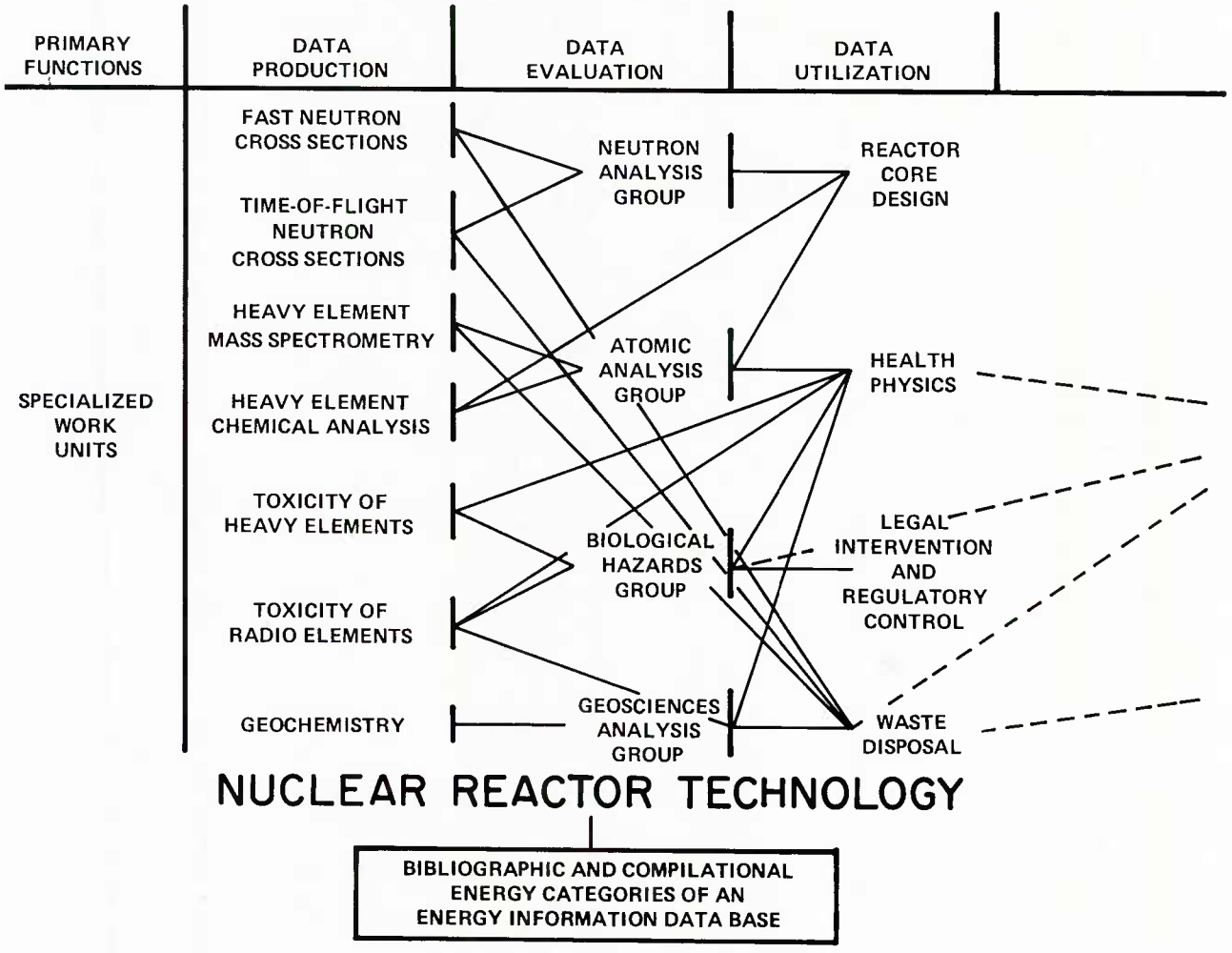


Fig. 5.0

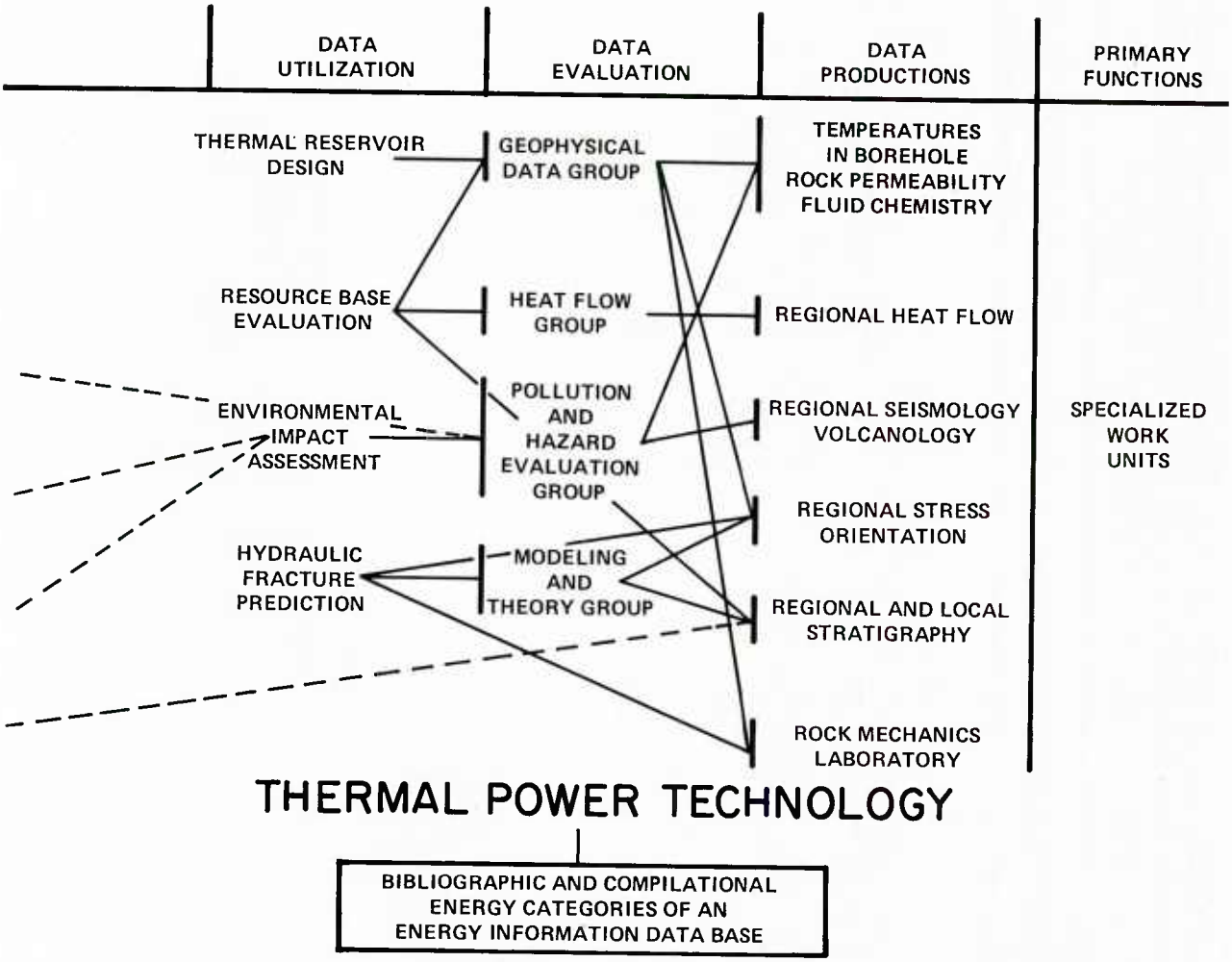


Fig. 5.1

Note that the bibliographic and compilational functions covering these activities are part of a much larger general energy data base.

This type of overview role could be performed by a properly constituted Data Analysis Center (DAC) and is so done in some cases, but more commonly the data centers have much more circumscribed responsibilities which restricts their usefulness and detracts from their optimum utilization.

To develop a favorable data cycle design it is necessary to examine the structures of the component functions of the data cycle as follows.

6. Program Definition and Management

In the following discussion it will be attempted to be general in citing organizational structures. For obvious reasons, one cannot expect a unique approach to the problems because of differences in national political administrative systems. At the same time the differences will not be so great that the management thread, at least, cannot be easily followed. For reasons of familiarity there will be a tendency to use the US federal system as a guideline.

In technical mission oriented departments of government, the mission can usually be identified by a single word, e.g. Energy, Defense, Environment, Transportation, and others. Programs (in divisions) are subunits usually obvious to mission accomplishment. The detailed management techniques vary both within and amongst departments for reasons generally not pertinent to the data activity. The effect of the differing management methods on optimization of data utilization is, however, of considerable importance. This arises because interdepartmental communication at the level of planning for data programs, data generation and handling is often poor or non-existent. This leads to redundancy of measurement, of procurement of expensive equipment and of dissemination of data to the user community. Since scientific and technical data are often, but not always, of much broader than departmental/divisional interest and content, it is of the utmost importance to make visible the data derived for departmental programs in order that other departments and agencies may make use of them as soon as possible. Program definition, direction and redirection is, of course, dependent on the applicable data bases available at every moment in time; these data bases may originate entirely within the department or more usually be a complex of data developed in-house, from other departments, from academia and industry.

In many mission-oriented organizations scientific and technology data production is done by means of a well defined management structure, but the evaluation and dissemination functions are likely to get very uneven treatment.

7. Data Production/Generation Organizations

The means by which data are produced under government department auspices will be discussed in terms of optimizing response to the needs of programmatic offices.

A large share of the numeric data of physics, chemistry, the geosciences and biosciences pertinent to departmental mission responsibilities are produced within an in-house laboratory complex as discussed earlier. The specific work items may be fully formulated by a program manager or may be submitted to program managers either through solicited or unsolicited proposals. Much of this work is highly original basic and applied research requiring complex and costly instruments and devices which are used to respond to a wide variety of problems. In a large department the support for these facilities may come from an applied research division which has a strong fundamental research orientation not dissimilar to that found in a large university, but other divisions with primarily applied science objectives may contribute to "requesting" measurements with various priorities established for a more narrowly focused and directed set of data needs of interest to themselves. Some facilities are highly unique or particularly versatile and may obtain support from various technical divisions inside and outside a particular department. The sources of the actual work will usually lie in large multi-program laboratories owned by the government and operated either by a contracting entity or by the government itself.

Such facilities with their associated scientific and technical staff are by far the most costly component of the data cycle, but, after all, they produce the information required to accomplish the conceived missions and objectives of this responsibility of the government. In multiprogram departments, having a well defined applied research division an optimum cost-effective approach to acquiring data desired by a multiplicity of program divisions is by the use of a cross cutting matrix management system. This provides the programmatic "user" divisions with a mechanism for obtaining the data needs from the use of the best facilities and their professional staff; it is a particularly desirable method when interests of more than one division in the same or related data from a particular facility appear. Although this type of management structure is well known and usefully applied in industry, it is not yet ingrained, or often not even acceptable, in organizations still heavily basic science oriented. This comes about because the area of concern here is at the interface between data producers who do non-directed research for users (often themselves) in a search for an understanding of nature, while we are mainly discussing data producers whose output has been created "on order" from users who have applied technical and engineering objectives to meet. This is not to say that programmatic users do not use their own facilities when possible; in particular they may have unique facilities which are not common in the area of disciplinary science, but are needed at high priority to satisfy problems of highly directed science and technology.

A dichotomy becomes apparent here, i.e. that the programmatic offices are major users, but simultaneously producers of some data. In the case of the disciplinary divisions represented as the producer column of the matrix, they are simultaneous users of some portion of their own data. This is natural and highly desirable from the point of view of morale; it should not interfere with a well managed data system although it often does so for various reasons. Each of the rows and columns of the matrix management system must have as a working member a fully qualified staff person from the facility or function represented.

It is important to remember that the functions represented in both rows and columns of the management matrix are never performed or implemented by the government administrative offices at the headquarters level, but rather by scientists, technologists and engineers who are the actual producers of data and the actual users. This means that good communication between users and producers is essential once measurement programs for a user have been approved and are underway.

8. The User Community

In some sense the user community is the most important component of the data cycle, especially when thought of in terms of an end-use economy. Although sometimes the user community is well known, it is in many cases elusive, particularly in those areas of industry and commerce where the utilization of quality data and information has not fully penetrated. The following broad types of users are well known.

- A user community may be entirely within a government department, for instance in military development or in an in-house space sciences program, but also solely for the purpose of elucidating national laws and phenomena.

- A user community may be entirely outside the public sector (in private industry or commerce) and may produce little or no data itself, but be a major user of data deriving from work funded by a government department for the same or totally different utilization. This community is usually fully eligible to receive such data and even create pressure to produce new data.

Another type of user may be a private industry acting as a direct contractor to a government department or be a joint-effort collaborator in a major applied project. And there are no doubt many other data users when specific cases are considered.

It is clear that the user community must be incorporated into the planning process for an efficient data cycle. In particular, requests for data usually originate with users who are not otherwise involved in the data production/evaluation steps at all. On the other hand some of the most sophisticated and necessary data are requested by producers and evaluators in order to improve their own capability to make difficult measurements or evaluations needed by the more applied users. This latter matter will become clearer when the necessity for standards, discrepancy analysis and new device instrumentation are discussed. A parallel statement can be made about developers of data bases since their requirements may reveal needs for data not yet produced.

Data bases are the fuel for activities primarily user oriented, i.e. they are designed to be responsive to specific needs of the user for his problem solving commitment. Thus the initial determination of specifications for a given user would best come from himself. Depending on the breadth of the component data requirements and the capabilities and training of the user, the data cycle may be entered at one or more points. The initial exploration is likely to be into the bibliographic/compilational functions which will then quickly provide guidance in to specialized areas of data assessment and evaluation. All data bases seem to require iteration and updating for completeness. In many cases the course of problem solving leads to tightened specifications especially with respect to limits of accuracy.

The user community associated with each set of applied data needs must have a management mechanism which allows it to communicate data requests effectively to data producers and evaluators. Each such user group must also establish its own internal priorities for the data requested and transmit this into a broadly based organizational structure which can incorporate request, priority and data specifications into the overall data cycle system to have the overriding priority established. This organization may also be able to help with scheduling when the work should be done.

9. The Bibliographic/Compilational Functions

Nearly eveyone who has engaged in original problem solving will be well acquainted with the necessity and importance of a bibliography of what has gone before in the problem area in question. In some sense, this is a starting factual data base in itself, and becomes even more so when the pertinent data in the references are extracted and compiled. From this most simple version of the functions being discussed above to the complexities of the National Technical Information Service, the National Library of Medicine, the Defense Technical Information Center in the USA, the World Data Centers for the Geophysical Disciplines, the Environmental Data and Information Service in the USA there is a wide variation in magnitude and degree of report and professional publication aquisition, storage in appropriate categories, and information and data retrieval. Much careful attention has been applied to the problems of extraction, storage and retrieval and the application of automatic information processing to this area makes it possible to anticipate that these essentially librarian functions will soon be under control. Some serious problems exist and are likely to remain for some time since they have to do with categories of scientific data and information which will require establishment of philosophies and policies by the scientific practitioners themselves. Such questions will no doubt be addressed by later lecturers.

A more serious problem for which the light at the end of the tunnel is not yet visible is the one of providing quality assured data now to be touched on.

10. The Data and Information Evaluation Function

A great deal will be said about this in subsequent lectures for several reasons. One of these is the fact that it is not always recognized how important this function is becoming even in non-technical uses. Another reason for the inadequate attention evaluation gets is that it must be done by scientists who themselves have experience in the subdisciplinary area they are evaluating which means that the manpower pool applicable is very small. A very serious management difficulty which besets evaluation work as a whole arises from the fact that there is no broadly established responsibility for the function nor is there even-handedness in the handling of this activity by departments or even within departments. This results in there being seriously inadequate funding to support the work which then falls behind and continues so and, even worse, there is no uniform support for evaluation even by the funders of the data producing function. There is thus no policy for the accomplishment of evaluations other than response to funding for special purpose assignments or at the desire of the evaluator himself for whatever his motivation might be. It would seem to be most desirable to establish at least discipline wide philosophy and policies for performing evaluations systematically.

Some attention should be given to the development of data bases from evaluated data in colloboration with the interested users. In this way the people most knowledgeable about the content of the data would be brought into the formation of the most useable output and could in fact also provide evaluation for data bases themselves.

11. Model of the Data Cycle

At this point it becomes desirable to present a more detailed data cycle than has been done above, since many of the reasons for the functions shown have now been discussed briefly. This model cycle is shown in Fig. 6. Some critical issues will be discussed in terms of the model and a crude management structure which is compatible will be given. It is very unlikely that a model even as incomplete as the one proposed can be put into use in a general way, even within a single government department. Nevertheless, in at least one division of a US department essentially all of the functions described have been accomplished usefully for many years; furthermore this particular program has been coupled into rather parallel international efforts with considerable success in establishing priorities, responding to requests from a wide variety of users, accomplishing data compilation and evaluation, and providing automated data storage and access. There is therefore some encouragement to be derived from this example that even wider networking may in time be feasible between not only subdisciplines, but wholly different disciplines. The necessity for such interrelatedness has become very apparent in data uses that depend heavily on data bases which cut across many scientific disciplines. An obvious example of this lies in the area of environmental science where physics, chemistry, biology, toxicology, the geosciences both in and above the earth become pertinent to problem solving in such diverse areas as chemical plant siting, hazardous waste disposal, the acid rain problem and the effect of CO_2 from fossil fuel power plants on long range climate.

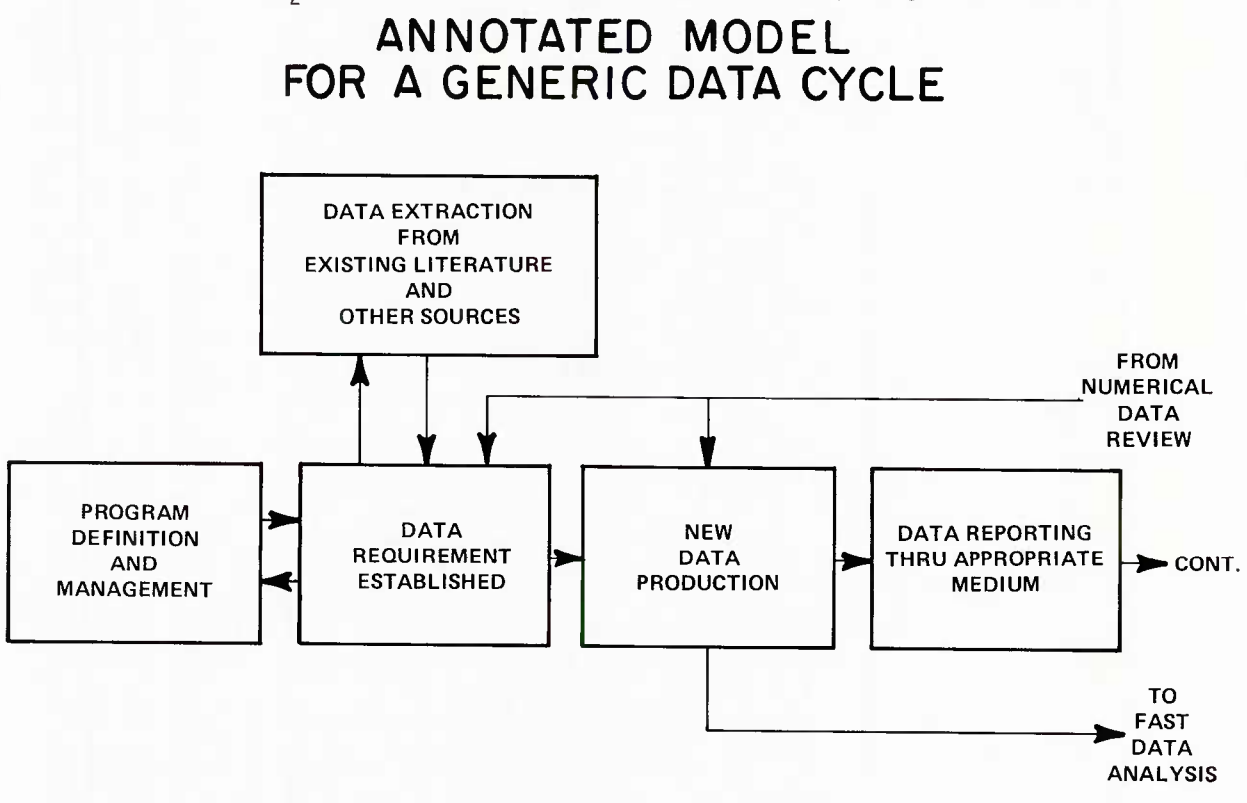


Fig. 6.0

With respect to Data Extraction when the Data Requirement is being established, a complete bibliography for the category must be made available identified as to quality of data and quality of evaluation.

With respect to Data Requirement, programmatic data requirement is generally in multiple subdisciplines. A single data production group is unlikely to be able to provide all the data for all the program needs.

With respect to data reporting, programatically "reporting" is done by:

- Publication in professional reviewed-journals-best
- Published as laboratory issued documents available to requestors-acceptable
- Published in classified documents even though only a portion of the data is classified-problem area
- Data remain in Laboratory notebooks in both raw and reduced form-inexcusable
- Data never brought to any useful professional conclusion-intolerable.

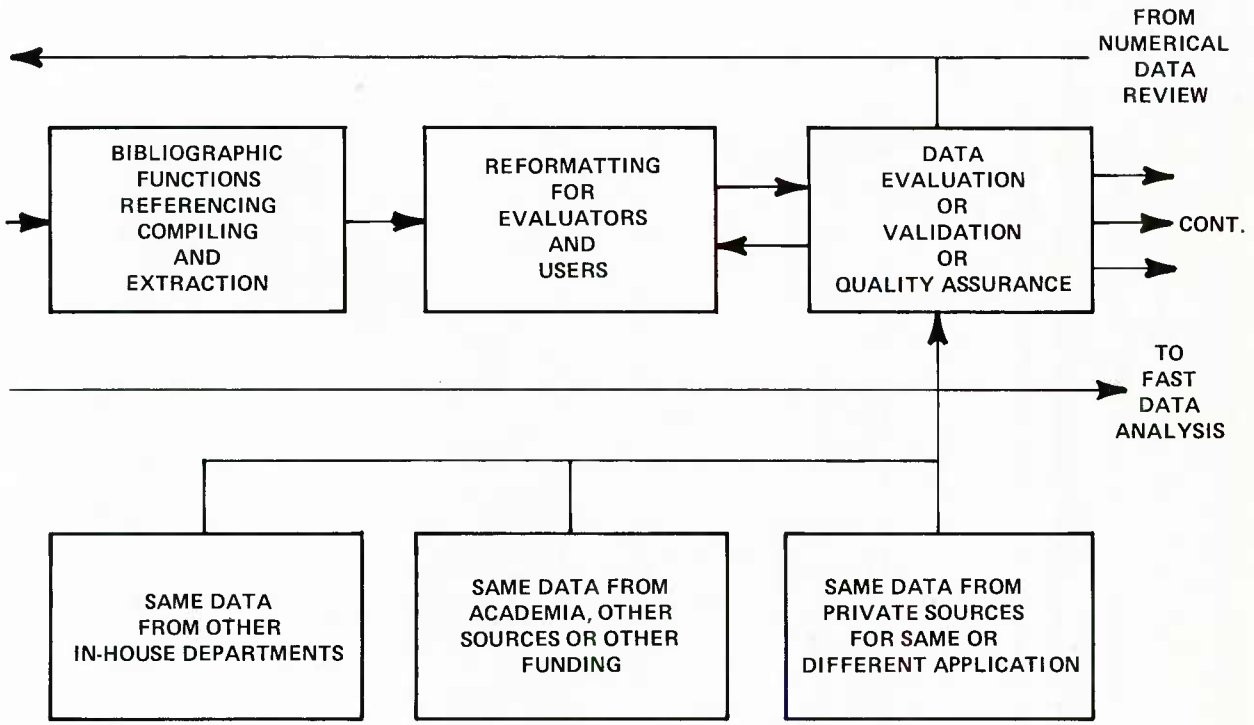


Fig. 6.1

With respect to Data Evaluation, a single evaluation group is likely to have expertise and experience of limited breadth, but in great depth of scientific and technical data.

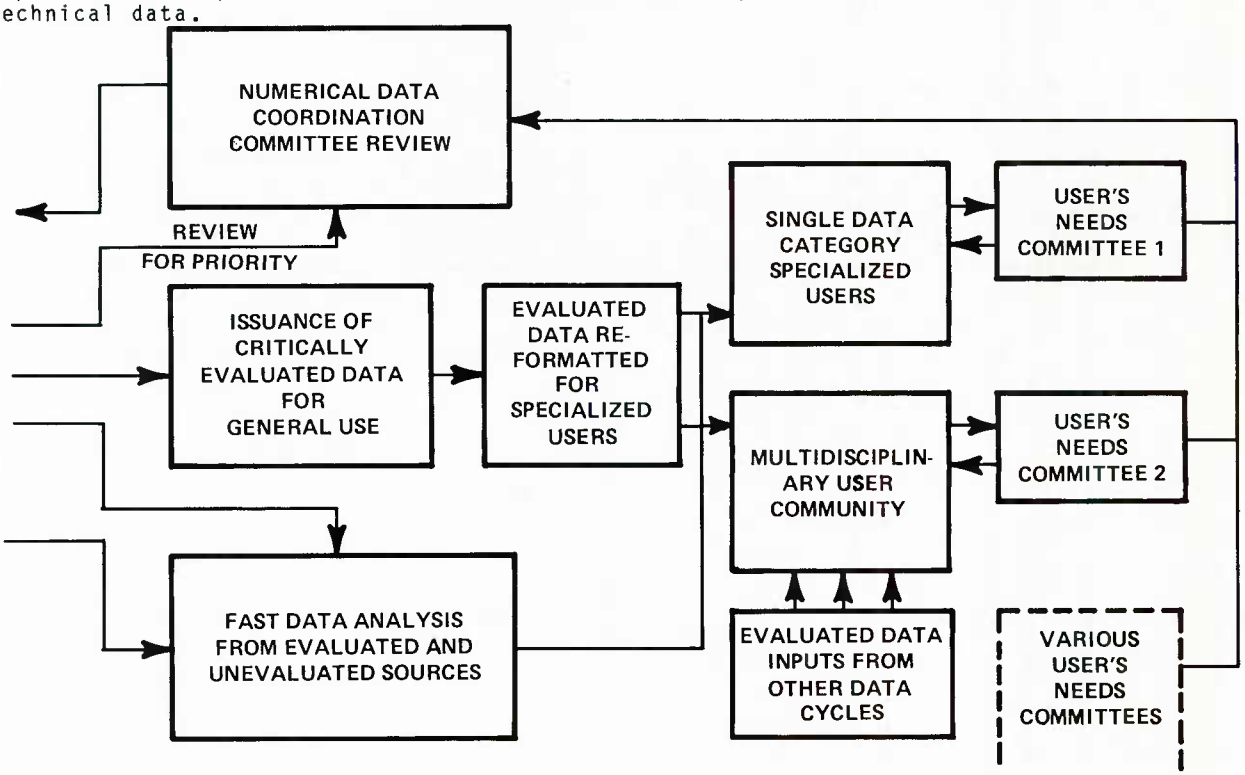


Fig. 6.2

With respect to Numerical Data, this important Committee must have subpanels of experts at least in such areas as:

- fundamental constants and standards production
- discrepancy analysis
- experimentalists with hands-on experience
- invention, design and construction of new measurement devices
- and others.

Neither model nor organizational system are so unique that the format shown needs to be closely followed; the model is shown primarily to serve as a guide to what can and what may work. In view of the fact that data and verifiable information is a, or even the, major output of those departments of government which are scientific and technical in nature, it is truly desirable to make a systematic and determined approach to improve what is done with the data generated. In fact, when this is not done, the data producing departments are abrogating their responsibility for their primary product.

12. Review and Advisory Management Structure for the Data Cycle

In order for the above model to work properly it is necessary that there be management subgroups within some of the functions shown which will arrange the implementation of necessary actions. For instance users external to a department which gets data produced must each have an internal-to-themselves data needs group to establish needs, priorities, accuracies desired, time required and other specifications. To accomplish this assignment the users must communicate with data evaluators and data producers; in a few situations this can be done directly, but more usually it will have to be done through a coordination committee.

Note that the data producers and data evaluators are either in, or are under the jurisdiction of, the department's division which responds to measurement requests coming from program divisions. For this reason each department program-division must also have a data needs committee or contract out that function. The requirements for these in-house needs are essentially the same as those of the external users with the additional responsibility of the in-house group to integrate all requests into a categorized request list.

An established channel must be set up for all data needs committees to transmit their requirements to the compilers of the request list and communicate with data producers and evaluators. This latter communication is for information purposes and carries no force of implementation. The official channel of communication between users and producers and evaluators of data could be a coordination group as follows.

13. Numerical Data Coordinating Committee

A full plenary group would be composed of representatives from each of the disciplinary subdivisions of a department's research division since this is where the program management responsibility is lodged. To these must be added appropriate representatives of those laboratory complexes where the data is actually produced either to order or by choice of the scientists. Appropriately chosen representatives of the evaluator community must also be on this committee.

This is now the decision making group which is in a single data category channel and

1) Accepts user input in the form of requests for data;

2) Melds requests of the same category and establishes an overall priority for each measurement item;

3) Transmits agreed upon requests to data producers and/or evaluators;

4) Monitors accomplishment of work;

5) Determines whether measuring methods and devices in existence are adequate to the requirement;

6) Sets up expert standing subcommittees on such matters as measurement status, standards, discrepancies, new facilities and new techniques;

Normally this committee would not meet in plenary session and would accomplish its assignment by having each of its meetings addressing only one or two specific data areas so that only those with the applicable generation, evaluations and utilization responsibilities would need to attend.

It must be apparent that the subject of data bases, the useful form of all the foregoing work has been only superficially mentioned. This is a result of the fact that the numbers and information which go into the data bases must first be produced, processed and quality assured. Unfortunately this may often take years to produce the inputs to a data base being formulated. The formulation of this last step is then complex, but no longer subject to the problems discussed above and will be given the expert attention deserved by later speakers.

CONCLUSION

What has been presented here is a large scale and even schematic map of what a fully implemented data cycle might look like and what might be needed to operate it. Following speakers, experts in various facets of this field will provide insights in

depth as to how the work is truly accomplished and what problems lie in the way.

My own thesis is that the total implementation of the utilization of data and information is of such great importance to our culture, nations and region that it must be accomplished in a far more explicit and organized way than now and in the past and will require careful study of how this can best be done. A second component of this thesis is that neither private nor public sectors of our political organizations can accomplish the longer range expected objectives alone thus requiring rational public/private arrangements to be established.

TYPES OF TECHNICAL CONTENT OF DATA

by
David G. Watson
Senior Research Associate
University Chemical Laboratory
Lensfield Road
Cambridge CB2 1EW
England

SUMMARY

The various uses of the word "data" are noted and a convenient definition of numeric data is proposed. Data can be broadly categorised in terms of the uses to which they are put or in terms of who uses them. The paper attempts to provide a brief overview of the characteristics of scientific data and does so with respect to a simple classification scheme. This scheme categorises data with respect to time, location, mode of generation, nature of the quantitative values, terms of expression and mode of presentation. It is illustrated by examples taken from the physical, earth and biological sciences. Finally some of the special problems associated with, for example, geophysical and biological data are noted.

It could be said that data constitute the life blood of human society. Each day we are presented with packages of data through the media of newspapers, radio and television. However there are only a few situations when an individual is in a position to check the authenticity of such data and the interpretations which have been placed upon them. As scientists and technologists, on the other hand, we have been trained to handle data in an objective manner and seek to apply the data without personal bias. It is probably true that the majority of scientists have, until recently, been concerned with data from the rather narrow fields of their own scientific specialisations. Increasingly, however, there is a need to utilise data from several scientific disciplines as part of a mission-oriented project and it is against this background that I wish to introduce some feeling for the great variety of scientific data types and some of the problems associated with them.

In the English language the word "data" is used rather freely by both scientists and non-scientists. A survey[1] of chemists, conducted some eleven years ago, revealed that 49% equated data with facts, 32% with numbers and 19% with other information. For most scientific disciplines it would be appropriate to say that data constitute the results of experiments or observations in the laboratory, workshop, field, observatory, clinic etc. For numeric data a convenient definition might be the following :-
A given piece of data normally refers to the magnitude of some quantity characterising some property or phenomenon of a certain system measured under a certain condition. Consider the boiling point of ethanol in the context of this definition. The quantity is the temperature, the property or phenomenon is the co-existence of liquid and gas phases in equilibrium, the system is the chemical compound ethanol and the measurement condition is standard atmospheric pressure.

The classification of data can be considered from a number of viewpoints. For example, in the technological field, one could group data in a broad fashion according to the category of person who needs to use the data. Thus the scientist or engineer will need property data, design data etc. for the purpose of research and problem-solving, whereas the manager and other administrative personnel will require product data, marketing data, regulatory and other legal data and so on.

Another facet of the data/user situation indicates three major groups of data :-

(i) Data which are generated in a specific discipline and used almost exclusively by specialists in the same discipline. Examples of such data would be crystal structure factors, seismographic records, electrocardiograms.

(ii) Data which are also used by scientists in a limited number of related disciplines, eg. data on ferromagnetic materials, steam tables, the genetic code.

(iii) Data which are used more widely,eg. the physico-chemical properties of organic and inorganic compounds, tide tables, toxicity data, human visual sensitivity to colours.

Some years ago[2] the Task Group on Accessibility and Dissemination of Data (a task group of CODATA - the Committee on Data for Science and Technology of the International Council of Scientific Unions) elaborated a simple classification scheme which is applicable to the broad range of scientific data. According to this scheme, data can be divided into six major categories :-

(A) Data with respect to the time factor
(B) Data with respect to the location factor
(C) Data with respect to the mode of generation or derivation
(D) Data with respect to the nature of the quantitative values
(E) Data with respect to the terms of expression
(F) Data with respect to the mode of presentation

(A) Data with respect to the time factor

Some data can be measured repeatedly, ie.they are time-independent (A1), whereas others can be measured only once, ie. they are time-dependent (A2). Most of the data of physics and chemistry belong to category A1. In astronomy, data on fixed stars would also be time-independent, whereas geophysical data measured on the occasion of an eclipse of the Sun would be time-dependent. In some cases where measurement is difficult, either for technical or financial reasons, then data which are, in principle, time-independent become time-dependent. Examples might include data from deep ocean beds or data collected by manned space vehicles.

(B) Data with respect to the location factor

Some data are independent of the location of the measured objects or phenomena (B1), whereas others are location-dependent (B2). As with the time factor, most data in physics and chemistry are location-independent. On the other hand, many data in the earth and astro-sciences are location-dependent, eg. data on rocks, fossils, meteorology. Equally, when ecological and biogeographical factors are considered then many biological data become location-dependent.

(C) Data with respect to the mode of generation or derivation

Primary data (C1) are data which are obtained by experiments or observations that have been designed specifically for the measurement of the particular quantities. Examples from chemistry, geophysics and biology are optical spectra, seismographic data and physiological data such as respiration rates, blood volumes etc.
Derived data (C2) are data derived by the combination of several primary data with the aid of a theoretical model. Well-known examples of derived data are the fundamental constants of physics and chemistry, the temperature distribution in the Sun and the genetic code. It should be noted, in passing, that the primary data used for the derivation of C2 data will usually have been processed to some degree, often referred to as data reduction. It is important that details of the data reduction process should have been fully documented so that the derivation process is based on valid primary data.
Theoretical data (C3) are data which are pproduced by theoretical calculations or predictions. Molecular properties calculated with the aid of quantum mechanics or the prediction of solar eclipses using celestial mechanics are examples of theoretical data. In genetics an example would be the prediction of the external expression of the hereditary constitution of an organism from the assortment of genes which are present.

(D) Data with respect to the nature of the quantitative values

Determinable data (D1) are data on a quantity which can be assumed to take a definite value under a given condition. Most of the macroscopic data of physics and chemistry are determinable, as are the elements of planetary orbits. In biology, with the exception of abnormal specimens, the chromosome numbers and gene loci are constant for a given species.
Stochastic data (D2) are data on a quantity which takes fluctuating values from one sample to another, from one measurement to another, etc., even under a given condition. Examples would be the shapes of polymers in solution, the structure-sensitive properties of solids, soil composition, solar flares and, in biology, most data on animals and plants of a given species are different from one specimen to another.

(E) Data with respect to the terms of expression

Quantitative data (E1) are measures of scientific quantities expressed in terms of well-defined units. Most of the data of physics and chemistry are quantitative, as are meteorological and physiological data. It should be noted that certain quantities are expressed by numbers using arbitrary scales. An example of such semi-quantitative data would be the Mohs scale of hardness.
Qualitative data (E2), taken broadly, may include any definitive statements concerning scientific objects, properties or phenomena. An obvious example from geology would be the description of rocks in terms of texture, colour etc.

(F) Data with respect to the mode of presentation

Numerical data (F1) are data presented as isolated numerical values and most quantitative data would correspond to this category.
Graphical or model data (F2) are data presented in graphical form or as models. This mode of presentation is commonly used to summarise data and to assist the user in his perception of the essential features of the system under study. Examples include phase diagrams, molecular models, geological maps, metabolic pathways.
Symbolic data (F3) are data presented in symbolic form, often for the sake of clarity and ease of understanding. An example would be the symbolic presentation of weather data.

The six major categories which have been described are not mutually exclusive but rather they provide a faceted classification scheme. Thus, for example, meteorological data can be regarded as time-dependent, location-dependent, derived, stochastic, quantitative, numerical or graphical.

Whereas the use of computerised bibliographic databases is now fairly routine, both by scientists and intermediary personnel such as librarians and information officers, the same cannot be said for numeric or factual scientific databases. Indeed, many people have expressed some unease concerning the "black box" retrieval of numbers and facts from computerised systems. The safeguards which must be applied will, no doubt, be discussed later in this series and probably amply illustrated with examples taken from the fields of chemistry, physics and engineering. I would therefore like to give a very brief account of some of the problems which are inherent in data from some other scientific disciplines.

In geophysics the observational data can be regarded as of two principal types[3], viz. monitoring or survey data. In the former the same parameters are observed repeatedly at a given location under conditions which change with time eg. data on volcanic activity or variations in the Earth's magnetic field. Surveys, on the other hand, involve repeated observations of the same parameter successively at different locations under conditions which do not change with time (at least for the duration of the observational experiment). An example of survey data would be measurements of water temperature at a specified depth along the path of travel of an ocean cruise. The proper interpretation of many data types in geophysics often requires that the scientist should have access to data on similar measurements made at other locations and, on an international scale, there is a clear need for systematic and timely exchange of data. In fact, during the International Geophysical Year in 1957, this need was met by the establishment of the network of World Data Centres. These centres, located in the USA, USSR, Japan and Western Europe, undertake to collect and exchange data in all the major branches of geophysics.

In geology the description of any stratum or rock body involves the recording of data on composition, shape, spatial orientation in relation to its neighbours, geographical location, age etc. It might have been expected that geologists would have developed a fairly high level of consistency in identifying the many different types of rocks which exist at the Earth's surface. In fact, international tests have been conducted whereby the same rock samples were examined in various laboratories and the resulting descriptions revealed considerable divergence from uniformity. Neverthelress, in recent years, some progress has been made to improve the consistency of data description. Mucch of the impetus for this has come from petroleum exploration activities where consistency in the recording of drill and core data is obviously of great importance.

The handling of biological data[4] introduces factors which are not present in the case of physical data. The three factors which distinguish living from non-living matter are reproduction, metabolism, growth and associated with each of these is variablity. In individuals variations can be produced by adaptation or acclimatisation. From generation to generation the variability can be further proliferated by mutations and selection, also by learning in the case of the higher animals. Another interesting source of variation is caused by the act of measurement itself. The use of drugs, anaesthetics etc. interferes with the normal biological functions and thus it is of paramount importance that biological data are recorded with very detailed descriptions of the environmental parameters.

In recent years political pressures have resulted in the establishment of a large number of databases concerned with environmental data. One such area is concerned with data on toxic substances in foodstuffs. The demand for these data can be attributed to a number of factors eg. the increased use of agrochemicals to increase food production, the growth of heavy industry resulting in food contamination by lead and other industrial pollutants, the increasing use, by the food industry itself, of dyestuffs, flavourings and preservatives. Even though governments have spent large sums of money for the establishment of appropriate databases it is quite likely that some of these are doomed to failure because the necessary data have not been recorded at the various stages of the long food production chain. Examples of such missing data might include the methods of packaging and storing the foodstuffs, the cooking techniques and the properties of the utensils, the nature of the water supply, the age of the plant or animal and the part of it which was used for analysis.

In conclusion, I would like to stress the point that, no matter from which subject area they are generated, data can be of optimal usefulness only if they are associated with some realistic estimates of their reliability. The critical evaluation of data with the assignment of quality indicators is one of the most important aspects of data handling and this topic will be developed in subsequent lectures.

REFERENCES

1. Slater,M., Osborn,A. and Presanis,A., Data and the Chemist, 1972, Aslib Occasional Publication No.10, London, Aslib.
2. Study on the Problems of Accessibility and Dissemination of Data for Science and Technology, 1975 CODATA Bulletin No.16, Paris, CODATA.
3. Shapley,A.H. and Tomlinson,R. in Data Handling for Science and Technology (ed. Rossmassler,S.A. and Watson,D.G.), Amsterdam, North-Holland Publishing Company, 1980, chapter 3B
4. Bartels,H. and Hausen,U. in ref.3, chapter 3A.

Sources Generating and Reporting Numerical and Factual Data

Dr. Malcolm W. Chase
The Dow Chemical Company
Thermal Group, 1707 Building
Midland, Michigan 48640 USA

Summary

The re-use of experimental data can be a frustrating endeavor. This is frequently due to the re-use of data in an attempt to solve problems for which the original experimental work was not directed. An experiment (or set of experiments) which is designed properly to solve a specific problem may well be severely lacking in sufficient information when applied to unrelated problems.

The nature of the laboratory (university, government, or industry) normally will often dictate whether the information is required for basic or applied technological purposes. Time, money and internal political pressures in turn affect the thoroughness of the study and the resulting publication. Often these pressures detract from the authors' awareness of the greater utility of their results. Thus minimal data can be even further limited as a result of the attitudes and publishing policies of the laboratory and journal. Examples of this interplay and its effects will be given in the areas of low temperature calorimetry and vapor pressure.

Introduction

This presentation will discuss the experiences of a laboratory in its never-ending search for "data". The Thermal Group of The Dow Chemical Company is a physical chemistry laboratory with primary strengths in thermodynamics. As such, the laboratory is involved in the quantitative aspects of chemistry. Our prime activities involve the experimental and theoretical/calculational determination of heats of chemical reaction, heat capacity and thermal conductivity of materials, vapor-liquid equilibrium in binary and ternary systems, and vapor pressures of chemicals as well as other thermodynamic/kinetic activities. At The Dow Chemical Company these data are typically used by chemical engineers to design safe and efficient production facilities.

In order to perform our activities efficiently, our experimental and theoretical/calculational efforts are designed to build on existing information. Thus a significant effort is directed towards being aware of and using available information which is pertinent to our physical chemical involvement. Literature surveys, technical meetings, and personal contacts form a solid basis for maintaining a high level awareness of relevant information. The type of data we need is typically used repeatedly. The thermodynamic/kinetic application is often a use of data in a scientific discipline for which the original work was not intended or anticipated. Thus our awareness must be broad and not restricted to the direct generation of thermodynamic/kinetic information. It must encompass any and all scientific disciplines from which relevant information may be derived. Of equal importance is the knowledge that pertinent information exists, that such information is retrievable, and that such information can be easily accessed time and time again. This talk will deal with the generation of such data and the reporting thereof, using thermodynamic examples for illustration.

JANAF Thermochemical Tables

Very early in the existence of The Dow Chemical Company, the research and production people recognized the utility of thermodynamic data to predict the feasibility of chemical reactions, equilibrium conditions, vapor pressures and the like. By 1940 it became apparent that the thermodynamic data culled from the literature was insufficient in both quality and quantity for the company's needs. At that time the Thermal Group was created to determine thermodynamic parameters and to evaluate the data from the literature.

The early systematic collection of data was accomplished under the direction of Dr. D. R. Stull. This early data collection and measurement was directed towards materials of company interest. A portion of this collected and determined data was published in 1956 under the title of "Thermodynamic Properties of the Elements" (Advances in Chemistry Series #18 by D. R. Stull and G. C. Sinke). The tabular format and the 298.15 K reference temperature used in this publication have been widely adopted by others working in the field. Coincident with the data collection, computer programs and techniques were specifically developed for the critical assessment of the data and the compilation of thermodynamic properties. Throughout this period, those people who evaluated literature values were also making measurements and determining the accuracy of the evolving methods so that they would have first-hand knowledge of the validity of the data. In many cases the literature data was validated by repeating the measurement in the laboratory.

In 1958 The Dow Chemical Company proposed a program for the United States government directed toward the development of high energy propellants. At that time the difference in various estimated heats of formation for compounds of interest could change the energy of a system by as much as 20%, so a strong thermodynamic section was

written into the proposal. This proposed work resulted in the JANAF Thermochemical Tables which are the oldest regularly maintained and updated compilation of temperature dependent thermodynamic properties of chemical species in the United States. Originally sponsored by the Department of Defense in 1958, they were later continued with Air Force funds. In 1976, the Department of Energy joined the sponsorship for a program of critical evaluation and compilation similar to that of the Air Force, but working with chemical species selected for their relevance to energy research. Currently, this project receives only Air Force support but the uses of the data have been extended far outside the bounds of propellant evaluation.

Thermodynamic data can be derived from a variety of different studies. Some of the more common sources are: cryogenic heat capacities, drop calorimetry; heats of combustion and reaction; equilibrium studies, vapor pressure, mass spectrometry, electrochemical cells; kinetic studies; matrix isolation spectroscopy, quadrupole electric deflection, electron diffraction, microwave spectroscopy, molecular beams; infrared and Raman spectroscopy, ultraviolet spectroscopy, and photoionization spectroscopy. This list serves as an indicator of the many diverse fields which provide pertinent data. New methods for generating such information are constantly being found, and each needs careful analysis for reliability, preferably by scientists who have first-hand knowledge of the methods, procedures and equipment used to make the measurement. This reflects the interdisciplinary uses of data (and techniques); the data generators are not always aware of all potential applications of their data. On the other hand, those in need of data may not fully appreciate the applicability of data from another discipline and its real contribution to their project.

The integration of these various data into a concise and consistent format is of great value to scientists and engineers, particularly those who are engaged in the development of propulsion systems and high temperature materials. Merely collecting and summarizing all of the thermodynamic data from the many diverse sources in one place is in itself a great value, and the critical evaluation and mathematical derivation of secondary parameters by those skilled in the art is an additional large benefit.

The JANAF Thermochemical Tables are documented such that each table contains the following information: (1) a tabulation of temperature dependent values (at 100 K intervals and at 298.15 K) of Cp°, S°, {H°-H°(298.15 K)},-{G°-H°(298.15 K)}/T, Δ_fH°, Δ_fG°, and log K, and (2) a text which describes the critical evaluation of the adopted values and itemizes input data used in the calculations, and the method of calculation used. Based on users comments, the previously accepted format of the text may be modified to include some additional information. For example, we now prominently display the dissociation energy for diatomic gases and the heat of atomization for polyatomic gases. In the future, we anticipate adding crystal structure data and more detailed numerical data at phase transitions. Since many users need to interpolate the values in our tables (perhaps using an equation), tabulated values at intervals less than 100 K would be useful for temperature ranges where the temperature-dependent functions change very rapidly and/or nonuniformly. Data users can greatly increase the value of data generated by others by actively commenting on additional data needs and the reasons/uses for such additional effort.

The quality and the useful range of each thermochemical table depend not only on the quality, quantity, and extent of the available experimental and theoretical data but also on the methods used to calculate the functions Cp°, S°, {H°-H°(298.15 K)}, -{G°-H°(298.15 K)}/T. In our study of chemical species, we decide on the degree of sophistication of treatment based on our interpretation of the quality, quantity and extent of available data. Existence of sufficiently extensive data for a diatomic gas, for example, may justify use of a direct summation treatment instead of the classical approximation of Mayer and Mayer in the statistical mechanical partition function. The more sophisticated treatment should give a significant increase in accuracy in the final product in order to justify the extra time and effort. We must judge from the available data whether this will be the case.

When the study of a particular species is completed, we compile a list of "missing data" which, if available, would significantly reduce the uncertainty in the thermochemical table. Our aim is to promote experimental and theoretical studies in those areas which are unknown or have a high uncertainty, instead of those areas which are well known. Additional studies should be directed at filling in the data gaps or confirming data which may be suspect.

In this summary of our activity with the JANAF Thermochemical Tables, four data generator concerns have been highlighted:

(1) the interdisciplinary uses of data;

(2) the communication of the data users' needs to the data generators

(3) awareness of the relationship between quality of data generated versus sophistication of the data application; and

(4) the usefulness of listings of missing or insufficient data.

These concerns will be mentioned, directly or indirectly, many times in this presentation.

Data Generating

In examining the data generated by any laboratory, it is oftentimes easy to be very critical of the quality, quantity, and extent of the data produced. How often have you questioned the thoroughness of a study, the quality of the results, the lack of understanding of the problem, the poor sample characterization, and so on. Although these (and others) are valid concerns, they should be tempered by an understanding of the project objectives under which the data was generated and the technical knowledge and experimental capabilities available at the time and at the location the project was performed.

Data needs are becoming more and more sophisticated with each passing year. Many years ago, say approximately twenty years, the so-called final results of many projects were data limited; the results were not limited by the data treatment techniques. Conversely, today there are more projects which are limited by the degree of sophistication of the data treatment technique employed, rather than by the data itself. Thus, the demands on the data generators are increasing. Data users are expecting and demanding an increase in the quality, quantity and extent of the data generated. This is a reasonable demand based on our current developing experimental capabilities and the increased technical understanding. For example, the automation of laboratory equipment certainly has removed many of the manual problems in the experimentation. More data points can often be recorded in a shorter period of time with more precision. This more easily obtainable data often reveals effects not previously observed. It is always appropriate to assess the quality, quantity, and extent of data generated in any project, but any criticism of the study itself should be made in terms of the knowledge at the time of the study rather than on today's knowledge.

The other aspect of the data generation problem concerns the project objectives. We try to be aware of the numerous existing laboratories which can produce data. These laboratories - national, private, and university - normally produce data as part of a defined project. In projects of interest to the JANAF staff, the thermodynamic data for a particular species, although a necessary portion of the project, may not be the prime activity. Herein lies a major frustration - the information of prime importance to the JANAF staff may not be of prime importance in the project producing such data. This frustration can be tempered somewhat by knowing the nature of the laboratory, the types of projects it is typically interested in, and the relative importance attributed to the various data produced. Such an understanding permits an appreciation of the whys and hows of a particular study as it was conducted. It can often provide insights as to where to obtain the data and how to persuade authors to expand their experimental study to be of more value to others. This could be especially easy if these multiple interests are financed by the same agency.

The quality, quantity, and extent of any data generated will be determined by the relative importance of such data to the overall project. This restriction is a result of the limited objectives and limited finances of the project. The data is generated to satisfy a particular end use, as such it may not adequately satisfy other end uses. Even in laboratories specializing in thermodynamic measurements, the amount and origin (which normally defines objectives) of the financial support will often impose greater restrictions on the quality, quantity, and extent of data than the nature of laboratory. Your project and its data needs may well not match the data needs of another project and as a result, the quality, quantity, and extent of data produced may not be mutually satisfactory for both projects.

Regardless of the priority of the many data generating activities in a project, the quality and extent must be carefully assessed. In the originating project the effect of the uncertainty in all generated data should be assessed as to its effect on the results of the overall project. The need for highly accurate/precise data may well in turn cause a reassessment on the time and financial commitments necessary for this work. On the other hand, if the effect of the data on the results of the overall project is such that a rather large uncertainty can be tolerated, a less precise experiment often may suffice.

In summary the data generating aspects of a laboratory are tied to the project objectives, which in turn are restricted by the origin and financial terms of the support. Such data can be used effectively provided an assessment as to the quality, quantity, and extent of the data is made. However, limitations in the quality, quantity, and extent of the data may result from the relative importance of the data to the overall study, the impact of the quality of the data on the final project results, the experimental capabilities available at the time of the project, and the actual end use for which the data was actually generated.

Data Availability

Data can arise from many laboratories - private, national and university. These laboratories have experimental and theoretical programs ranging from basic to applied research. Although these aspects certainly affect the data generating activities, the

larger effect arises from the amount and source of funding. It is the author's contention that the project selection is determined from the type (private, national, or university) and nature (basic or applied) of the laboratory. These laboratories apply for funding in project areas which fit their goals and objectives. Of course, these are tempered by the rise and fall of the presumed importance of various disciplines. There is certainly a tendency to go with the easier money or the more fashionable research of the day. As mentioned earlier, the relative importance of any specific data to the overall project and the end use of such data in that project will dictate the quality, quantity and extent of the data generated.

In the development of any project, the investigators themselves must perform the role of a data evaluator first and the role of a data generator second. The investigators, in appraising the overall data needs of their project, should be aware of all the available relevant literature. Then the proposed data generation activities should build on this available information. The investigator must access all the available information. This can be viewed as a three-step process. First, can a scientist verify that the data does, or does not, exist? Second, knowing that the data exists, can the scientist obtain the data? Third, having obtained the data, is all the essential information available? Compared with the data generating activity itself, this three step procedure is not only inexpensive but also requires little time. However, it is not always performed efficiently or completely, if at all.

Does the data exist? The ease with which this question is answered depends on the attitudes and publishing policies of the previous data generators. The more a data generator makes use of the available data, the more a data generator is aware of the data recovery problems. Hopefully each successive data generator will improve the ease of data recovery.

Each scientific discipline normally has a handful of journals (or the like) which specialize in this discipline. If the study is published in one of these journals which indeed does specialize in the same field of activity, then most scientists verify the existence of the data by manually scanning this journal. If the study is published in an unrelated journal, then such data may not be found easily. Of course if the material is not published at all, one can imagine the difficulties in knowing such data exists; such knowledge rests solely on personal contacts. The investigator must be aware of the normal location of the data of interest so as to search in the best locations.

The expansion of the number of journals and, more importantly, the expansion of the interdisciplinary use of data renders a solely manual search inefficient. A more encompassing survey can be made through the use of the abstracting services. These abstracting services are numerous and include Chemical Abstracts, Physics Abstracts, and National Technical Information Services. The key assumptions here are that the study is indeed published and made accessible by the choice of the proper key words for the abstractors. In addition, it is important to choose a title that properly describes the study and to state in the abstract the data that are really available in the article. The authors and abstractors play a vital role in assuring that all data in a published study is indeed recognized and cited by the abstracting services. It is not unusual for secondary data to be buried in a publication and be rendered essentially inaccessible due to a lack of a visible reference to it in the abstract, for example.

There are many abstracting services. Although there is much duplication in coverage between them, each refers to some journals/documents not covered by the others. Thus by using the abstracting services, either manually or by computer, the scientist can cover a much larger number of journals quickly. Two words of caution, however, are necessary. First, always use more than one abstracting service (especially true for thermodynamics). Second, each abstracting service has its own strengths and weakness - learn what they are. For example, in the use of Chemical Abstracts, if you are interested in obtaining all the data on the vapor pressure of boron, you can search under vapor pressure and boron. Interestingly, you will find more vapor pressure data for boron when searching under "boron" than if you searched under "vapor pressure". Of course, if the journal chosen for publication is not covered by the abstracting services, the data is rendered inaccessible by these sources. Another not to be forgotten source of possible data is the governmental research directories and the like.

A specialized spinoff from the abstracting services is the bibliographies which contain listings of more specific data. For example, in thermodynamics, there are the Bulletin of Chemical Thermodynamics, Bibliography of High Temperature Chemistry, and Index Thermochimique. These publications give references to the available data. Although these contain mostly information also available from the abstracting services, they often have additional sources. The main advantage is, of course, the collection in one place of all the specialized data. In addition, there may well exist comprehensive compilations which summarize and/or critique the available data.

Can the data be obtained? Local libraries and interlibrary loan programs normally permit access to the data which is published in the more common journals. The use of obscure journals and data depositories may slow the recovery of the data but at least the data is available. Proceedings of conferences, government documents, and national

laboratory reports may be difficult for the scientist to obtain directly, but skilled librarians can normally obtain this latter information easily. Again, the assumption is that such reporting is available to the abstracting services.

Is all the pertinent information available? Any proposed data generation should build on existing information. One of the many advantages in obtaining previous publications of related work is to gain better insight into the potential experimental problems. Duplication of previous studies may be necessary for verification or calibration, but the investigator must ensure that the same experiments are being performed and that the experimentation is correct. The failure to recognize earlier related studies will undoubtedly reduce the value of the new study.

Data Reporting

The preparation of a high quality comprehensive publication requires considerable time. An in-depth description of the study, including all data generating activities, may not be possible under the funding constraints of the project. Funding agencies are seemingly encouraging publication in a respected journal with a sound reviewing policy. This is their way of ensuring the quality of the research results. Limitations in the publication may well match the relative importance of the various data generating aspects of the project.

The data reporting activities may be classed as published and unpublished. Keep in mind that primary and secondary data generated in a project may not truly belong in the same journal. Consider the following questions:

A. Where are the results published:

   1. in a journal consistent with primary project objectives

   2. in a series of unrelated journals, since prime data generation and secondry data generation differ

   3. as quarterly/annual progress reports

   4. as laboratory reports (available outside laboratory)

   5. as a talk at a technical meeting (proceeds published)

B. If the results are not published, are they available:

   1. as laboratory reports (not readily available outside of laboratory)

   2. as confidential/restricted information

   3. by letter, telephone

   4. announced in compilations/bibliographies/meetings

   5. in certain data depositories

In this series of questions, normally the published categories are accessed by the abstracting services, whereas the unpublished categories are not. Thus, the second category leads to frustration in the data recovery process. Publication in an obscure, hard-to-obtain journal or certain data depositories or in an unabstracted contract summary will often render the data inaccessible.

Given the fact that the researcher has decided to publish the results, the publication will exhibit two types of limitations in the degree of transfer of information. The publication will be author-sponsor limited and publisher limited.

The author-sponsor limitations center on:

1. considering the entire project, the more important data generating activities in regards to its primary result will receive more emphasis

2. the importance of publication to sponsor and author (will determine location of publication and completeness of documentation)

3. awareness of other possible uses of data.

The publisher limitations center on:

1. space limitations (no experimental data tabulations?)

2. philosophy of journal (type of data information)

3. review policies (too strict or not strict enough).

There is undoubtedly a tendency to restrict publication to the new results without addressing the comparisons with existing information or the effect of secondary

information on the prime result. A tabulation of the experimental data, along with a graphical display of such in comparison with other data, is invaluable in any article. Tabular comparisons are not as informative. If equations are also necessary, they too add to the value of an article, but having actual data available is most important. The time required to construct a high quality article and the time required to reach publication are often two independent problems. A brief report may result as a compromise. There are occasions, however, when an in-depth article will require a substantial rewrite through the review process. Time or desire not permitting such activity, the article is resubmitted and published in a less-critical journal. The quality of such a publication is greatly diminished. Also consider that some progress reports on laboratory reports may undergo varying degrees of review. The data users need to be aware of such processes, so as to properly assess value of work of the data generators.

Data Needs

In order to efficiently resolve a project's objectives, the investigators should build on existing information. New data generation should add to the available information and be consistent with the quality required for the project's end use. A decision as to making new measurements and/or using existing data must be made intelligently. If no relevant data exists, it may not be clear as to why such a study has not been done (no interest, too difficult, etc.). On the other hand, if relevant data does exist, it may not be easy for the investigator to judge its value for the project of interest. How many scientists do you know who have their own favorite data sources? At the same time do you think that they are aware of the currentness of this data source, let alone the effect of its uncertainty on their project? The data generators often do not make sufficient use of the available data. Nor do they make use of other scientists who may well be very familiar with the existing data and its problems.

There are three types of data which may be needed:

1. data for which no similar data exists

2. duplicating a data source for verification, preferably using a different technique

3. multiple data sources exist, a definitive study is needed to reduce significantly the uncertainty of current data.

In effect these three types deal with reducing uncertainties, with (1) referring to an establishment of verified data and (3) being a fine tuning of the data. This latter type study (when necessary) is more difficult to justify financially as it involves fixing the "so-called 3rd decimal place or the like" and is not easily appreciated by those in control of the purse-strings. The difficulty lies in making the researcher aware of actual data needs and how these related to the project of interest. A scientific advertising campaign is needed. Perhaps a compilation of data needs, complete with probable experimental difficulties, is needed.

Vapor Pressure of Boron

There are over twenty reported studies which relate to the heat of formation and vapor pressure of boron. These studies involve the sublimation of boron and the decomposition of boron carbide. Four of these articles will be used to illustrate the generating and reporting of numerical and factual data. Each article will not be discussed in detail; only certain aspects pertinent to this talk will be discussed. These studies are:

1. "Mass Spectrometric Study of the Vaporization of the Titanium-Boron System," P. O. Schissel and O. C. Trulson, J. Phys. Chem. 66, 1492-6 (1962); manuscript received March 16, 1962.

2. "Mass Spectrometric Determination of the Heat of Sublimation of Boron and of the Dissociation Energy of B_2," G. Verhaegen and J. Drowart, J. Chem. Phys. 37, 1367-8 (1962); manuscript received April 30, 1962.

3. "The Sublimation of Boron," R. W. Mar and R. G. Bedford, High Temp. Sci. 8, 365-76 (1976); manuscript received September 1, 1976.

4. "Phase Relationships and Thermodynamic Properties of Transition Metal Borides. 1. The Molybdenum-Boron System and Elemental Boron," E. Storms and B. Mueller, J. Phys. Chem. 81, 318-24 (1977); manuscript received June 18, 1976, revised manuscript received November 30, 1976.

As background information, the CODATA Task Group on Key Values for Thermodynamics critically reviewed all the data on the heat of formation of gaseous boron which was available for the mid-1970's. They recommended values for the heat of formation and its uncertainty of gaseous boron, $\Delta fH°(B, g, 298.15\ K) = 560 \pm 12$ kJ mol^{-1}. Since the uncertainty was thought to be too large, additional investigations were desirable. The latter two articles mentioned above had not been published at the time of the CODATA review. Mar and Bedford do not recommend a value but report values of $\Delta fH°(B, g, 298.15\ K) = 563.6 \pm 33.4$ and 561.1 ± 3.3 kJ mol^{-1} from torsion effusion data and $\Delta fH°(B, g, 298.15\ K) = 566.1 \pm 14.3$ kJ mol^{-1} from Langmuir data. Storms and Mueller reported

$\Delta fH°(B, g, 298.15 K) = 574.9 \pm 0.8$ kJ mol^{-1} from Knudsen effusion data. It is disconcerting that the value reported by Storms and Mueller, rather than confirming the CODATA recommended value, lies outside the range of the CODATA recommended uncertainty.

1. Schissel and Trulson

Schissel and Trulson, with financial support from the Advanced Research Projects Agency, studied the vaporization of the titanium-boron system. At the time of this study, the use of boron compounds in high energy fuels had created a need for thermochemical and thermodynamic data for such compounds and their combustion products, including those products caused by the reaction with materials of construction. The work by Schissel and Trulson at Union Carbide Corporation was part of this effort.

These authors, citing only four previous sublimation studies, stated that there was agreement in the reported heat of sublimation values. Thus, it is not clear as to the rationale for studying the sublimation of boron. Four sets of experiments were conducted. For each set a heat of sublimation was derived at an "average mid-range" temperature. The auxiliary data, used to convert the measured data to a heat of sublimation value at zero kelvin, was stated. The authors stated that these results are in excellent agreement with existing data and thus proceeded to use existing data rather than data from this study. In a four page article, the entire boron sublimation study is relegated to a short paragraph.

The sublimation of boron was obviously not the prime motive for this study. The authors directed their publication effort to providing information on the vaporization in the titanium-boron system. The boron information provided is useful but much of its value is lost because of the minimal information provided. The number of data points in the four sets of experiments is unknown. More importantly, the data points are not available in this publication. Even with the information provided, two major questions remain: what experiment was actually performed and what is the definition of an "average mid-range" temperature.

Two final comments need to be made concerning this publication. First, the abstract contains no information as to the inclusion in this work of a study of the sublimation of boron. Second, there are available from the laboratory of Schissel and Trulson many government progress reports which include experimental data for the sublimation of boron. Is there a relationship between this data and that alluded to in the article in the Journal of Physical Chemistry? The published article is easily accessible but the government progress reports are not.

2. Verhaegen and Drowart

The work by Verhaegen and Drowart, performed with the financial support of the U.S. Air Force, was published as a Communication in the Letters to the Editor section of the Journal of Chemical Physics. As with the work of Schissel and Trulson it was part of the early propulsion effort directed at understanding boron chemistry. This work was performed at the Free University of Brussels in Belgium. Verhaegen and Drowart, in studying the boron-carbon system, needed a reliable value for the heat of sublimation of boron for subsequent thermodynamic computations. Concluding that eight previously reported values were discordant, the authors decided another determination was necessary. Recall that the Schissel and Trulson study referred to only four previously reported values, stated that the values were in agreement and yet submitted their article for publication within two months of the study by Verhaegen and Drowart.

The emphasis in this published study is the heat of sublimation of boron. The values resulting from this mass spectrometric study were tabulated in comparison with results from other studies. In other words, there was a table of heat sublimation values at zero kelvin. The actual experimental data, converted to vapor pressure, was presented only in a small graph, log p versus 1/T. The auxiliary data used to derive the heat of sublimation values were stated. Corrections to the values reported by Verhaegen and Drowart can be made as improved auxiliary data becomes available.

With the main emphasis of this work being directed towards the knowledge of the heat of sublimation of boron, the vapor pressure values were of secondary importance and were not tabulated. The vapor pressure values could be obtained by digitizing the graphical information. The graph, as published, was 5.5 x 8.0 cm. By the way, one would not normally expect secondary information in a Letters to the Editor section of a journal.

After publication of this article, Professor Drowart generously made available the experimental data to the JANAF staff. As a result a thorough analysis of the data and the detailed comparison with other studies was now possible. Not only could a graph of all vapor pressure data be constructed for visual assessment of the agreement (or lack thereof) of many studies, but also a detailed statistical assessment of the data and its reduction to a heat of sublimation could be made.

In summary the data of prime importance for the JANAF staff is the actual vapor pressure measurements. This information was not of prime importance in the article published by Verhaegen and Drowart and thus was relegated to a small graph. The value of work was enhanced greatly by the authors' willingness to make the experimental data publically available. Keep in mind that the primary interest for Verhaegen and Drowart was the mass spectrometric investigation of the boron-carbon system which was published

in detail in the Journal of Chemical Physics, but not in the Letters to the Editor section. In this boron-carbon system publication, there is only reference to the boron heat of sublimation value. The heat of sublimation for boron was important to this study but was not the prime goal. These authors chose to report this work separately and briefly. A similar situation existed in the study of Schissel and Trulson, but these authors relegated the boron sublimation to a small portion of the article.

3. Mar and Bedford

The boron sublimation study of Mar (Sandia Laboratories) and Bedford (Lawrence Livermore Laboratory) is intended to be a definitive study. The source of funding for this study is not specified within the publication. The authors stated that a reinvestigation of the vaporization behavior of boron was undertaken because of the large uncertainty associated with the value of the heat of sublimation. Vapor pressures were determined by torsion effusion and mass spectrometric techniques. In addition, the authors were concerned with establishing the evaporation coefficient of boron.

The awareness of previous studies and the possible rationales for the cause of the discrepancies is an important aspect of this study. Mar and Bedford not only summarize the results of the previous studies but also direct their experimentation to resolve the presumed causes for the discrepancies. Without belaboring the issue, this article provides essentially all the information needed by the JANAF staff. At this time it is immaterial to discuss the quality of this study. The point is that sufficient information is provided to evaluate the quality of this study and the effect of its results on the previous studies. Also, the results appear to support the recommendedations of the CODATA Task Group.

4. Storms and Mueller

Storms and Mueller, at the Los Alamos Laboratory of the University of California, studied the molybdenum-boron system as part of a program to evaluate the possible usefulness of molybdenum borides in thermionic energy conversion and magnetohydrodynamic generators. The thermodynamic understanding of this binary system requires knowledge of the heat of sublimation of boron. Storms and Mueller felt that the large uncertainty in this value made a new study necessary. They referred to the CODATA Task Group review of the boron data as well as the same previous studies quoted by Mar and Bedford.

The main emphasis is in the binary system although the authors devote much effort to the sublimation of boron itself. Such activity is reflected in the title of the article and in the abstract. Unfortunately 136 data points are summarized by thirteen heat of sublimation values. A data dispository or the like was not used. However, the authors were well aware of the previous studies and their problems. The present study not only uses improved procedures but also offers a rationale to explain the discrepancies (different from those suggested by Mar and Bedford).

It is interesting to note that the studies by Schissel and Trulson, Verhaegen and Drowart, and Storms and Mueller had as their primary interest a binary system. They all had need for the heat of sublimation of boron. Although they all actually published minimal results, the amount of information transferred and the relative importance in the publication varied significantly. In all cases, however, regardless of type and nature of the laboratory and funding, the minimal information is a result of the secondary nature of the boron sublimation to the overall project.

Heat Capacity of Boron

As a second example to illustrate the generating and reporting of numerical and factual data, the heat capacity and enthalpy studies of β-rhombohedral boron are interesting. Of the twenty-two studies, only the three more recent low temperature studies will be discussed. Suffice it to say that similar problems exist with the high temperature enthalpy studies. The three studies to be discussed are:

1. "Low Temperature Heat Capacities of Inorganic Solids. V. The Heat Capacity of Pure Elementary Boron in Both Amorphous and Crystalline Conditions Between 13 and 305 K. Some Free Enèrgies of Formation," H. L. Johnston, H. N. Hersh, and E. C. Kerr, J. Amer. Chem. Soc. 73, 1112-1117 (1951); Chem. Abstr. 45, 6036a.

2. "Temperature Dependence of the Specific Heat of β-rhombohedral Boron," V. I. Bogdanov, Yu. Kh. Vekilov, G. V. Tsagareishvili, and I. M. Zhgenti, Fizika Tverdogo Tela 12, 3333-3336 (1970), Chem. Abstr. 74, 5758.

3. "Low Temperature Heat Capacities of Open-Structured Crystals," N. Bilir, Dissertation, Department of Materials Science and Engineering, Stanford University, 1974; Chem. Abstr. 82, 117008t; Diss. Abstr. Int. B 35(6), 2963-4 (1974).

The JANAF staff is interested in critically evaluating all heat capacity and enthalpy studies for β-rhombohedral boron. Sufficient information is hopefully provided in each published study so that the quality and extent of the data can be assessed not only within each study but also in combination with other related studies. These three

studies just mentioned specifically involve the measurement of the heat capacity and are the studies from which the best low temperature thermodynamic description of β-rhombohedral boron can be derived. The three articles vary significantly in the amount of information transferred to the reader. All three articles were easily available, both in terms of finding the reference (through Chemical Abstracts) and in obtaining a copy of the published study.

1. Johnston, Hersh and Kerr

The study by Johnston, et al. at the Ohio State University was supported partially by the U.S. Office of Naval Research. It was part of an overall program to measure low temperature heat capacities of inorganic solids. This particular study involved the measurement of 46 data points in the temperature range 16.90 to 303.71 K. This publication contains the following information:

- specific calorimeter used and a reference to the operation and construction thereof.
- method of sample preparation.
- "d" values from X-ray diffraction.
- relative atomic mass value for boron (10.82)
- tables of 46 data pairs; temperature and heat capacity
- graph of data; including data points and smoothed heat capacity curve.
- smoothed table of derived thermal functions

The objectives of this study and the objectives of the JANAF evaluation match extremely well. Thus this article presents most of the information deemed important by the JANAF staff.

The experimental data was presented in two forms - graphical and tabular. This permits later investigators to easily reuse and reanalyze this study. In particular it permits an easy comparison with other related data as it becomes available. In addition, the authors calculated smoothed values of the thermodynamic functions at 25 K intervals (and at 17 K and 298.16 K) by graphical integration of the heat capacity curve.

Rather than discuss all aspects of the evaluation process, only four items will be emphasized. First, the type of calorimeter used and the expected errors as a function of temperature are important information necessary to assess the quality of the resulting data. This information is provided by Johnston, et al. In addition, the performance of the calorimeter on other samples (or standard materials, if possible) is relevant. This information is often not given by authors (at least, not directly) and may be difficult to obtain. In the case of the calorimeter used by Johnston, et al., readers must read numerous articles published on other materials to gain an appreciation of the quality of their work.

Second, the description of the auxiliary data used in the study is necessary so that, if the auxiliary data changes through the years, the results can be corrected. In this study, such a datum is the relative atomic mass of boron.

Third, the authors discussed an unexpected result of their study. The data revealed a shallow maximum at ~25 K in the heat capacity versus temperature graph. A rationale was presented to explain the behavior. Existing studies which revealed similar behavior were also discussed.

Fourth, although the authors describe extensively possible methods of preparing samples of pure boron and then detail the actual method used, the major deficiency in the study is a lack of a definitive purity analysis and the identification of the specific crystalline modification. Later investigators claim that Johnston et al. actually studied the γ-modification. Fortunately, enough information was presented in the publication so as to judge the possible characterization of the sample. The deficiency, so to speak, is not in the published article, it is in the study itself. The deficiency became apparent only years after the study was completed. It is now known that many different crystalline modifications exist for boron.

2. Bogdanov, Vekilov, Tsagareishvili and Zhgenti

A similar study was reported 20 years later by Bogdanov, et al. These authors measured 30 data points in the temperature range 16 to 280 K. This work was conducted at the Moscow Institute of Steel and Alloys. It appears that, although heat capacity measurements were made, the main emphasis is the phonon description and semi-conductor behavior of β-rhombohedral boron. This article contains the following information:

- specific calorimeter used and a reference to the operation and construction thereof.
- method of preparation of β-rhombohedral boron.
- X-ray diffraction results for unit cell parameters.
- purity determination for metals.
- relative atomic mass value for boron (10.81)
- graph of θ_D vs. T (data points and smoothed curve)
- smoothed table of derived thermal functions

The actual measured heat capacity values are not given, either in tabular or graphical form. For the JANAF staff to obtain the heat capacity values, the graph of θ_D vs. T must be digitized and the heat capacity values back-calculated. The heat capacity is proportional to $(T/\theta_D)^3$. The error introduced by this process must be recognized in contrast to the experimental error. The published graph is 5.5 x 3.5 cm.

Following the same four items as discussed for the published study by Johnston, et al., numerous additional comments can be made. The type of calorimeter used and the probable errors are discussed. The performance of this calorimeter on other materials is not given. The auxiliary datum, i.e. the relative atomic mass of boron, is given and is different from that used by Johnston, et al. The two data sets cannot be compared directly. The data of Johnston, et al. must be converted using the proper (i.e., more recent) auxiliary data. The reported data did not support the results of Johnston, et al. in that a maximum in the heat capacity at 25 K was not observed. Bogdanov, et al. attributed the maximum in the heat capacity to impurities in the sample used by Johnston, et al. Sample preparation and the resulting purity is discussed but is still not definitive. The authors, however, state that this study reports the first measurements of the heat capacity of the β-rhombohedral modification of boron.

3. Bilir

In a third study by Biler in 1974 at Stanford University, again the heat capacities at very low temperatures, 2-20 K, were measured. The emphasis was on the properties of open-structured crystals with boron being one of many materials studied. This dissertation contained the following information:

- specific calorimeter used and construction thereof.
- origin of sample, stated to be β-rhombohedral boron.
- purity determination for metals.
- graph of C/T vs. T^2.
- graph of C/T^3 vs. T^2.

Although this dissertation and a corresponding talk were easily accessible (at least in the United States), the heat capacity values can be derived only from back-calculation of digitized results from the graphs. Many of the same comments made on the Bogdanov, et al. study apply here also. These comments will not be repeated. Two additional remarks are necessary. Telephone calls and letters resulted in a futile attempt to obtain the experimental data. Biler stated that his values mesh with those of Johnston, et al., and no reference is made to the values of Bogdanov, et al.

The data provided by these three studies permit the JANAF staff to generate a reliable thermochemical table for boron. However, the data is not of sufficiently high quality to require a highly sophisticated linear regression analysis. A fairly low level treatment involving graphical and mathematical techniques will be sufficient.

In retrospect, the absence of the tabulation of the experimental heat capacity data in two of these three studies is frustrating. Nevertheless the availability of the data in all three cases would not significantly reduce the uncertainty in the thermochemical description of boron. More importantly, the characterization of the sample in all studies is a serious limitation. In three studies mentioned here all were easily accessible but only one transferred all the data actually available in a easily reuseable form. Again the quantity of data transferred appears to be related to the main emphasis of project rather than the type and nature of the laboratory.

Conclusions

The value of any data generating study is dependent not only on the capability of the investigators but also their awareness of any relevant studies. The proposed study should build on existing information, being fully aware of its strengths and weaknesses. New work should attempt to dispel as many of the potential/real problems of the previous studies as possible and truly add to the value of the existing data. A detailed explanation of the procedures followed permit subsequent investigators to ensure the correctness and thoroughness of the study. A presentation of the results in the form of tabulated raw data and graphical comparisons with relevant earlier studies increase the value and the extent of re-using published work. Subsequent use of equations or the like is dependent on current and perhaps future needs. The extensive re-use of the results, especially in interdisciplinary uses, will be dependent on the awareness by the author of possible other uses, the completeness of the study, and the information transferred.

The generators of data should be concerned that the quality, quantity, and extent of the data needed or generated is consistent with the objectives of their project. The actual data produced (quality and quantity) may be tempered by the availability of existing data. As far as quality is concerned, the generators should be fully aware of the effect of data uncertainties in their project. For example, if the design of a heat exchanger is unchanged by thermal conductivity data of ±10%, then the generators need not spend extra time and money to measure thermal conductivity to ±1%. However, if higher precision data is readily obtainable with little or no extra effort, then the generators have a choice to make, but would hopefully go with the higher quality data.

The availability of data and the thoroughness of the study is dependent on many items. In today's activity, the project objectives and financial bounds impose greater restrictions than the type and nature of laboratory. All funding organizations want high quality results in their prime objectives of the project for their increasingly tighter money supplies. In addition, the emphasis in government funding is placed on publishing the data in a reviewed journal as proof of the quality.

All work done is dependent on the scope of the overall project. Certain aspects may be reduced in priority as to the quality, quantity, and extent of data produced in order to fit into the financial requirements of project. Support data (information of secondary importance) is often not given sufficient effort in order to "save" money. The acquiring of secondary data of sufficient accuracy/precision commensurate with previous goals is not always recognized.

Given that certain data exist, how much is transferred to the reader. The mode of publication and the degree of information transferred again depend on the relative importance of this information in the overall project and the financial constraints. The journals impose further restrictions as to the amount and form of information to be included. However, the authors and sponsors could do much to ensure transferral of all information by a different journal selection, by use of data depositories, and by pressuring journal editors as to the need of publishing all pertinent information.

The contention in this lecture is that the nature and type of laboratory dictates project activity. The data generating aspects are then restricted by the objectives and finances of the project to the extent that a priority is assigned to each data generating activity. This priority governs the quality, quantity, and extent of all data generated. This priority should be directly related to the effect of the uncertainty of such data to the overall project.

You, as a data user, may see limitations in this data due to (1) the interdisciplinary use of the data (a different end use), (2) lack of interplay between users and generators, (3) lack of awareness of the requirements of data quality and data end use, and (4) lack of awareness of related studies.

Extraction and Compilation of Numerical and Factual Data

J.H. Westbrook
Materials Information Services
General Electric Research and Development Center
Schenectady, NY 12305 USA

Extraction and compilation of data embrace the processes of achieving awareness of the existence of data, gaining access to the data, extraction of data from its source, knowledge organization and the compilation procedures preliminary to the critical evaluation of data. Each of these processes in turn subsumes a number of individual steps. In reviewing the subject, cognizance is taken of the individual needs and viewpoints of both generators and users of data. Changing norms of practices, work-habits and technology are also having significant impact on extraction and compilation procedures.

INTRODUCTION

The importance of the extraction and compilation of data to the progress of science and technology can hardly be overestimated. Such compilations realize several different objectives: convenience of access, formats structured for particular application, and condensation and homogenization of voluminous and heterogeneous observations. Once in hand, data compilations convey still other benefits: perception of unsuspected patterns, detection of errors and gaps in the data, and a basis for the testing of theories. It is therefore not surprising that the collection of astronomical data by Ptolemy remained a key reference for a millennium, that the first organized pharmacopoeia was still judged worthy of printing for the first time 1400 years after Dioscorides wrote it and that the systematic arrangement of the elements by their atomic weights by Mendeleev over 100 years ago continues to yield fresh insights in chemistry, metallurgy and physics.

In many fields, for example: motor vehicle registrations, stock market transactions, or census data, building of files of numeric data is quite straightforward compared to the case of technical data. The data need not be extracted from some other context, they are already reasonably homogeneous as a result of their method of acquisition, and they are likely already compiled in some structured fashion. Today they may even be digitized if derived from a word processor, photo-typesetter or analogous electronic device. In contrast, data in the technical field are usually buried in the primary literature, contract reports or other sources. Their diverse origins and the lack of extensive standardization of reporting contribute to a heterogeneity that impedes ready comparison. Such difficulties and idiosyncracies could be, and largely were, tolerated a generation or so ago; when the volume of data generated was but a tiny fraction of that now amassed each year; when technical data in any field were of concern to but a few hundred scientists and engineers rather than critical to a broad geographic, disciplinary, and application dispersion of all human endeavors; and when computers had, as yet, made no demands on data homogeneity and structuring.

Today we cannot afford a casual and laissez faire attitude to what has become a central issue in technical, economic, and political progress. The choice of subject for this lecture series amply attests to this view as does the initiation of various cooperative, trans-disciplinary activities in the field of data compilation and evaluation. Among the most notable of these we can list the establishment of the Tables Annuele de Constantes et Données Numeriques in 1910, the Landolt-Bornstein Tabellen in 1883, the International Critical Tables project by the International Union of Pure and Applied Chemistry in 1919, the formation of the Committee on Data for Science and Technology (CODATA) by the International Council of Scientific Unions in 1966, and the Committee on Engineering Information of the World Federation of Engineering Organizations in 1969.

It is significant that all of these are cooperative, trans-disciplinary activities. This circumstance reflects the recognition that the interests of both generators and users of data are involved. Such cooperation is essential, not only to share the enormous amount of work to be done, but perhaps more importantly to effect the necessary standardization of data reporting and analysis, to establish priorities for data extraction and compilation, and to allow for feedback from data users as to the adequacies of all steps in the process. The influence of computerized techniques for collecting, indexing, analyzing, and disseminating data permeates everywhere - easing many mechanistic problems but raising new ones of format standardization, system compatability, etc.

In the present paper we shall treat the extraction and compilation of data as embracing the processes of achieving awareness of the existence of data, gaining access to data, extraction of data from its source, knowledge organization and the compilation procedures preliminary to the critical evaluation of data. As will be seen, each of these processes in turn subsumes a number of individual steps.

AWARENESS OF THE EXISTENCE OF DATA

Abstract Files

Before one can extract, compile or evaluate data one must first become aware of their existence. This problem has been exacerbated by the exponential growth in the volume of technical information generated as illustrated by the famous plot of Price(1) reproduced in Fig. 1. The advent of on-line searching of computerized bibliographic abstract files [a 1979 estimate was $\sim 4 \times 10^6$ searches per year(2)] has ameliorated but not solved the problem. Price judges that this technique contributes no more than a factor of 10 improvement in locating relevant information whereas a factor of 300 is needed.

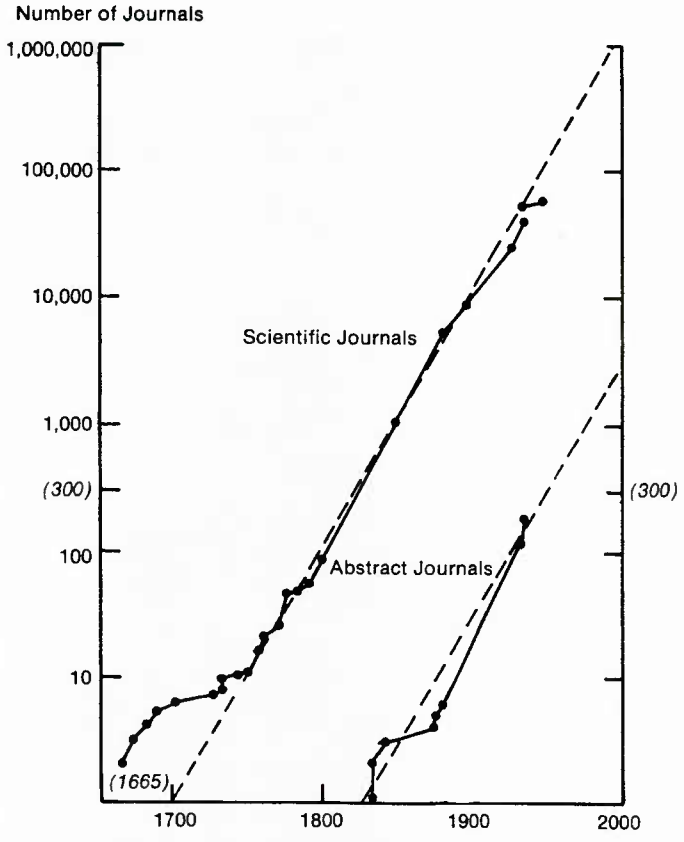


Fig. 1. Number of journals founded (not surviving) as a function of date [after Price(1)].

While the on-line bibliographic services do include some databases outside the regular journal literature (e.g., the National Technical Information Service abstracts of technical reports, ~70,000 items per year) a large fraction of potentially useful data remains unindexed and hence unknown to the majority of potential users. Private files, for example, are enormous, are still growing and proliferating. The DuPont Co. has a file of detailed information on about 100,000 substances of interest to it(3) and Union Carbide lists 130 properties of 1000 compounds.(4) General Electric maintains a materials information system with about 10,000 pages of properties information on materials used in its production activities.(5)

Reviews

Review articles complement the abstract services in attempting to cope with the problem of maintaining awareness of accumulating knowledge and information. It has been estimated that review articles constitute 3-4% of the total literature and each review typically cites 70 or more references.(6) Reviews come into existence by a variety of mechanisms. Authors are sometimes commissioned by their employers or some contracting agency to prepare a critical review of work in a designated field. Other times they will do so at their own volition as a result of a need to digest and structure the information as an adjunct to their own technical work. Finally they may be encouraged to do so by some editor, aware of their competence and possible interest. While all these review activities are a positive benefit, taken together they are still fragmentary and uncoordinated. Certain topics are reviewed every two or three years in one medium or another, while other topics can go for decades without a comprehensive critical review. Despite these shortcomings a great impetus to technical review activity has come about through the advent of regular review serials in a variety of fields. The complementarity and similarity of growth of this medium to the abstract collections is shown in Fig. 2 which plots the data on review serials listed by Chen(7) superposed on the data from Price in Fig. 1. Even though this listing is incomplete, the parallelism to the growth of journals and abstract collections is inescapable. Another lack is the absence of a comprehensive index to review articles although some indices exist to reviews in a limited field, e.g., organic chemistry.(8) Some of the standard abstract series, e.g., Chemical Abstracts, now provide a coded indication to characterize review articles, and ISI began in 1974 publication of an Index to Scientific Reviews. Variable definition as to what constitutes a review article, lack of distinction between critical and descriptive reviews and failure to achieve true comprehensiveness limit the value of these efforts.

Data Indexing

Even if abstracts and reviews gave adequate entrée to the general literature in a field, this would not suffice as a guide to the whereabouts of useful data. What is needed is a guide or index to data themselves. Data indexing may be either coincident (simultaneous with original publication or abstract) or retrospective. Coincident data indexing is held to begin in 1967 with a Canadian recommendation to index geological and other resource data on a national basis.(9) Subsequent efforts in this field, largely by Lerner of A.I.P., have led to the concepts of "data flags" and "data tags."(10) A data flag is an indicator of the presence of numerical data in an article or abstract. A data tag is a more descriptive indexing which characterizes in some depth the data reported. Murdock(11) has reviewed the subsequent efforts by IUPAC, NSF, CODATA, and INIS (IAEA) to encourage data indexing of all kinds. The recommendations of CODATA are

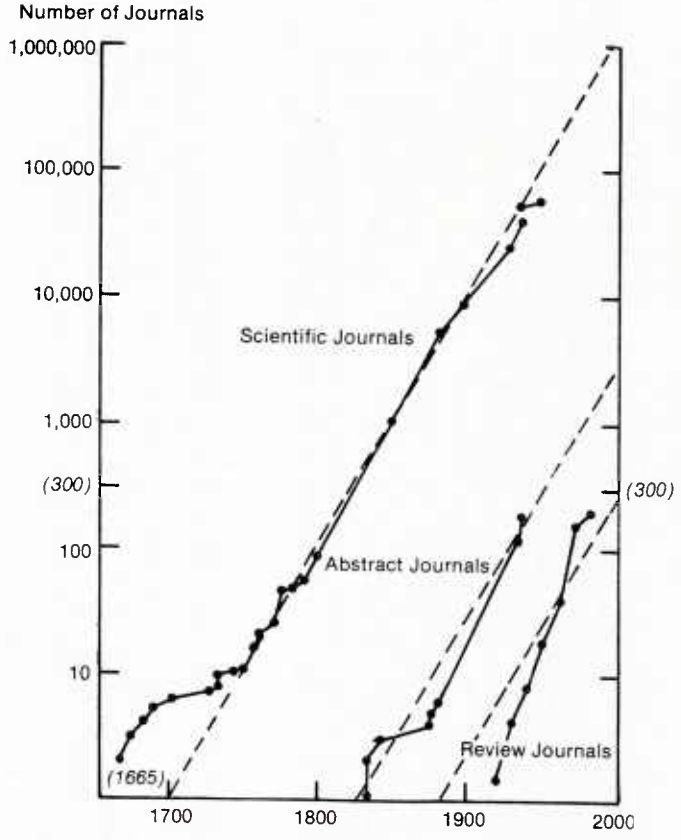


Fig. 2. Number of journals founded (not surviving) as a function of date. The plot of Price(1) is supplemented by data of Chen(7) for review journals.

contained in their Bulletins #12 and #19.(12)(13) As a minimum, a data summary should include a narrative statement that qualifies the data appearing in the original article, the major variables, the mode of presentation of the data (single points, tabulated values, graphics or parametric summarization), the characteristics of the data (measured, predicted, interpolated or extrapolated, etc.) and the degree of reliability. Major factors impeding further progress in this area are the costs of adding the additional depth and facets of indexing to present abstracts and the absence of significant standardization in this field.

Retrospective data indexing gives the seeker of data a very large assist by the compilation and publication of directories of data sources for science and technology. Data sources in this context refer not to the primary literature but to publications, information centers or on-line files which offer compilations of data taken from various primary publications, collected, homogenized as to units, description etc., arranged in a uniform format and (possibly) subjected to a degree of evaluation. Bell Laboratories(14) produced a KWIC index to the Annotated List of Data Compilations in the library of the Office of Standard Reference Data of the National Bureau of Standards. Wood(15) similarly has analyzed by subject field two editions of the directory of databases and computer products published by NTIS. Regional approaches to data guides include "Data Activities in Britain"(16) and "Guide to European Sources of Technical Information."(17) A guide to sources of geophysical and solar data has been published periodically by the International Council of Scientific Unions' Panel on World Data Centers (WDC's).(18) This covers the WDC's themselves, data handooks, data catalogs and instruction manuals for the presentation and exchange of data.

A first attempt at a comprehensive international directory was CODATA's "International Compendium of Numerical Data Projects," published in 1969.(19) Later it was decided to broaden the scope of this activity, to make it more detailed and systematic and to provide for more frequent updating. In the new scheme, CODATA is now publishing the "CODATA Directory of Data Sources for Science and Technology" as a series of individual chapters, each chapter constituting an entire issue of the CODATA Bulletin. The chapters published to date are as follows:

Bulletin	#24	June 1977	Crystallography
	35	Dec 1979	Hydrology
	36	Jan 1980	Astronomy
	38	Sept 1980	Zoology
	42	June 1981	Seismology
	43	July 1981	Chemical Kinetics
	48	June 1982	Nuclear and Elementary Particle Physics
	49	July 1982	Atomic and Molecular Spectroscopy

A related project sponsored by UNESCO and CODATA resulted in another book, "Inventory of Data Sources in Science and Technology - A Preliminary Survey"(20), which contained these chapters:

General data sources for science and technology
Renewable energy resources
Fertilizers
Hydrological sciences and water resources
Nutrition
Pesticides
Soil Science

Quite separately, for one of the Annual Reviews series, Westbrook and Desai(21) produced an article, "Data Sources for Materials Scientists and Engineers," which had much the same objectives and organization as the CODATA publications previously cited.

Another type of guide to data is being developed by a CODATA Task Group headed by Schoenberg.(22) The motivation of this task group is to identify gaps in the data on industrial organic chemicals by compiling an index to their property listings in standard reference works. For the purpose of this project industrial organic chemicals are

taken to comprise those produced in the U.S. at over 100,000 lbs/yr or selling for less than $1/lb, well-known hazardous or toxic chemicals, "fine chemicals," and a few selected others. Pure hydrocarbons (containing only H and C) are excluded on the basis that data sources for these are relatively well known. It is estimated that industrial organics so defined comprise about 4000 individuals. The tabulations being compiled show for each compound the chemical name, the structural formula, the Chemical Abstracts Registry No., the Chemical Abstracts name, and synonyms. For each compound the volume and page listing for about 30 chemical and physical properties in each of approximately 40 standard reference works are provided. Upon completion, data searchers will have available a very convenient directory to individual numeric data. Perhaps more importantly, gaps will be highlighted where significant data on industrially important organics have not been compiled (although perhaps measured and available somewhere in the primary literature).

Data indices discussed to this point, whether coincident or retrospective, have been primarily in print. There has, understandably, been pressure to move toward the availability of data indices on line. We cite here a few examples of on-line data indices. In 1982 CRC Press offered for the first time on-line access to the merged subject indices of over 1000 professional level reference books in science and engineering from several prominent publishers via a new product they call "Superindex." The American Society for Metals and SDC, Inc. have collaborated to produce Metals Datafile, the numeric equivalent to the Metadex bibliographic abstract file on metals and alloys. This new service is in part retrospective (covering leading handbooks, reference works and data collections in the metals field) and in part coincident (adding specific data to currently produced abstracts). The on-line aspect and the expanded search capabilities are attractive, but Metals Datafile has not yet achieved a broad and enthusiastic response from the materials community because of its use of unevaluated data, the arbitrariness of the selection of data for inclusion, and the incompleteness of its coverage. Improvements can undoubtedly be expected in the future. Hampel et al(23) and Merrill and Austin(24) have described an on-line master index of more than 4000 databases (both bibliographic and numeric) and models in the energy and environmentally-related fields. Another on-line directory to energy, environmental and socioeconomic databases and software programs is operated at Oak Ridge National Laboratory.(25) The U.S. Geological Survey has implemented an on-line system including the Water Data Sources Directory and the Master Water Data Index.(26) The National Oceanic and Atmospheric Administration operates the Environmental Data Index, a composite of several referral databases to 6000 individual environmental numeric databases.(27) Oak Ridge National Laboratory also has, both in print and on-line, a "National Inventory of Biological Monitoring Programs.(28)

Data Journals and Depositories

Other aids to readers to facilitate the awareness of data are data journals and data depositories. An outstanding example in the first category is the J of Physical and Chemical Reference Data, initiated in 1972 and published quarterly by the American Chemical Society and the American Institute of Physics for the National Bureau of Standards. This journal is not intended for the publication of primary research nor for review articles of a descriptive or primarily theoretical nature. Rather, it aims to provide collections of critically evaluated physical and chemical property data, fully documented as to the original sources and the criteria used for evaluation. Critical reviews of measurement techniques whose aim is to assess the accuracy of available data in a given area are also included. A topical analysis of the 215 articles published to date in this journal appears in Table 1. Although the categorizations are arbitrary and admittedly somewhat overlapping, a heavy concentration on thermodynamics and atomic and molecular data is nonetheless evident. Macroscale properties and behaviors of complex systems seem to have been slighted.

Table 1

Topical Coverage in J. Phys. Chem. Reference Data

	No. of articles
Thermodynamic Data	49
Molecules in Space	21
Atomic Energy Levels	20
Chemical Data Relevant to Air Pollution	11
Viscosity and Thermal Conductivity of Fluids	11
Molecular Vibration Frequencies	10
Atomic Spectra	10
Molten Salts	8
Properties of Water	8
Molecular Spectra	7
Electrical Properties	7
Liquid-Liquid and Liquid-Vapor Equilibrium	6
Diffusion	5
Atomic Transition Probabilities	5
High Pressure Data	4
Thermal Properties	4
Reaction Kinetics	4
Refractive Index	3
Atomic Form Factors and Scattering Cross-Section	3
Elastic Properties of Alloys	3
Miscellaneous Physics Data	10
Miscellaneous Chemical Data	6

Other journals devoted primarily to the recording, analysis, manipulation and projection of data as opposed to their interpretation and discussion include:

J. Chemical Documentation (1961-
J. Chemical and Engineering Data (1956-
Calphad J. (1977-
Physik Daten/Physics Data (1975-
Phase Diagram Bulletin (1980-
World Power Data (1949-
Atomic Data and Nuclear Data Tables (1969-
Organic Magnetic Resonance (1969-
Int. Jnl. of Chemical Kinettics (1968-
J. Chemical Thermodynamics (1969-
Mass Spectrometry Bulletin (1966-

Other continuing publication series partake of aspects of both the abstract journals and the data collections. These reference works extract newly reported data from primary sources and republish it with just such descriptive information as is required for minimal interpretation. The data are frequently brought to a standard format but are only rarely evaluated. Sources of this kind include:

Wohlbier, F.H. ed, "Mechanical Properties - Materials Reference Series 11: Trans Tech SA, Trans Tech House, CH-4711 Aedermannsdorf, Switzerland; Trans Tech Publications, 411 Longbeach Parkway, Bay Village, OH 44140. Issued in two or more volumes per year, beginning in 1975 (Vol. 1). Expanded to include corrosion properties in 1979 and since 1982 split into an applied and a fundamental series, "Key Engineering Materials" A1 (1982) and Single Crystal Properties" B1 (1982).

Wohlbier, F.H., ed. "Diffusion and Defect Data - Materials Reference Series 1," Trans Tech SA, Trans Tech House, CH-4711 Aedermannsdorf, Switzerland; Trans Tech Publications, 411 Longbeach Parkway Bay Village, OH 44140. Issued in two volumes per year, beginning in 1974 (Vol. 8). Continues Diffusion Data which began in 1967.

Stevens, J.G. and Stevens, V., "Mossbauer Effect Data Index," IFI, Plenum, New York, 1966 (annual).

"Structure Reports" Utrecht: Bonn, Scheltema, and Holkema 1913 (1928 annual).

"Selected Data on Mixtures", International Data Series, 1975 ___,

Data depositories are intended for archival storage of data, too extensive or of too specialized interest for inclusion in primary journals. Among these are:

American Chemical Society (Microfilm Depository Service), 1155 16th Street, Washington, DC 20036.

National Auxiliary Publications Service, American Society for Information Science c/o Microfiche Publications, 440 Park Ave. South, New York, NY 10016.

American Institute of Physics (Physics Auxiliary Publication Service) 335 E45th Street, New York, NY 10017.

National Research Council of Canada (Depository of Unpublished Data, C15T1), Ottawa, Canada K1A 052.

Unfortunately these files are not usually well indexed and the only key to the existence of a particular data set frequently lies in a footnote provided in the original research publication.

Information Analysis Centers

A valuable aid that has arisen in recent years is the socalled information analysis or data analysis center. These are physical locations with expert staff in some selected subject field or fields. The field may be discipline or mission oriented or may focus on certain large scale phenomena. These centers may be supported by some government agency (the National Bureau of Standards through its National Standard Reference Data Program manages about a dozen) or a trade association (e.g. the Copper Data Center at Battelle Memorial Institute. While these centers may perform some library-like functions of archiving and bibliographic referral, more importantly they make available to the inquirer their expertise in a limited subject field to direct him to still other experts or to provide him with data which has been evaluated, manipulated or reformatted. Despite the obvious utility of information analysis centers and the small cost (a few tenths of a percent of the cost of original data generation), their managers and supporters have not generally been successful in quantifying these cost-benefit ratios. As a result many centers have closed or operate at a much reduced level. Others, apparently needed, are never established. Carroll and Maskewitz provide a definitive review(29) covering the period through 1980.

On-line Data Files

On-line or (more broadly) machine-readable data files are the newest medium for numeric data access and one that is burgeoning rapidly. Several attempts have been made in recent years to catalog such files, the most notable being those by Williams(30) and Cuadra Associates.(31) Williams restricts herself to only bibliographic databases. While the Directory produced by Cuadra Associates includes numeric databases as well, the majority of these lie outside the scientific and technical field and deal with financial, business,

socioeconomic and similar fields. Wanger and Landau(32) in reviewing the Cuadra Directory found only 12 of 400 non-bibliographic databases referred exclusively to properties data in science and technology. Luedke et al(33), Luedke(34), Fried(35) and Tomberg(36) have also reviewed recent progress in this field. The problem of achieving awareness of the existence of specific data of interest when stored in this form is still unsolved. Plans to build integrated systems of related databases have been described by Heller and Milne(37) and by Westbrook and Rumble(38). Another approach is an on-line system with special "front-end" programs which would automatically direct the potential user to that database or databases which would best serve his particular need.(39)(40)

Data Needs

All of the efforts described above, while useful and generally praiseworthy, have still not solved the problem of "Is property x of substance y known, and if so where may it be located?" Indeed more than one reputable scientist has asserted that today is it quicker and cheaper to go into the laboratory and remeasure the value in question than to attempt to find it in the literature or another source. While, the past decade or so have seen the establishment of various data compilation and analysis projects, they are still too few to do a comprehensive job over all science and technology. Data users therefore have an obligation to prioritize their needs and publicize these findings. Examples of the latter are CODATA's publication on data needs in the energy field(41) and the American Chemical Society's study of the data needs of chemists.(42) A more generalized expression of data needs is found in a 1978 National Academy of Science report.(43)

GAINING ACCESS TO DATA

Awareness of the existence of data of possible interest is only a first step, various barriers may still have to be overcome before direct access is obtained. Among these are:

language
national borders
copyright
other proprietary considerations
unpublished character
military classification
"need to know"
transient availability

Language

The problem of language is a minimal one for several reasons. An increasing proportion of the world's original literature is in English, for those journals not in English many cover-to-cover translation journals exist, and English language abstracts may give enough idea of the data content of an article to judge whether a custom translation is warranted.

Transborder Data Flow

The United States is at once both the world's leading importer of data and the dominant exporter of information as shown schematically in Fig. 3.(44) This state of affairs has not been viewed with equanimity by many other nations who sense the possibilities of loss of independence and increased economic and social disadvantages.(45) Various actions have therefore been taken or proposed by individual nations to restrict transborder data flow. These actions can have both substantial economic effects (especially on the socalled transnational corporations), but also in the present context can inhibit access to scientific and technical data. Remote sensing technology, another dimension of change in the information field, is seen by some as a blatant invasion of national sovereignty. Recently, considerable apprehension has arisen in the U.S. about leaks to the Eastern bloc countries of data and technical information of military significance either via the open literature or commercially available services. This concern has led both to a National Academy of Science study(46) and to the use by the Defense Technical Information Center of a warning statement from the International Traffic in Arms Regulations on all DTIC database products.

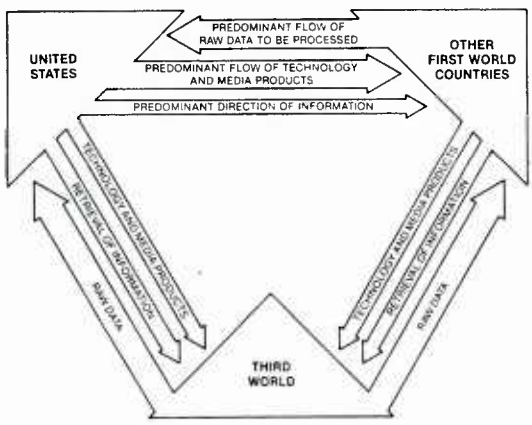


Fig. 3. Schematic of world-wide transborder data flows.(44)

Copyright

Attention is increasingly being given to copyright issues as regards the international accessibility and dissemination of numeric scientific data. Both the implications for use and access to published works and the implications for computer-readable works (including on-line) are involved. The whole topic is very much under study at the present time. Keplinger(47) reviewed the topic recently in ARIST, the CODATA Task Group on the Accessibility and Dissemination of Data has issued a memorandum by Fivozinsky(48) and King Research has published a study(49) instigated by the International Federation of Library Associations and Institutions focussed on machine-readable bibliographic records. It is concluded that the problem is as yet unresolved but that prompt, objective and foresighted actions must be taken or undesired restrictions on data access and exchange will result.

Other Proprietary Considerations

Much technical data lie in the hands of private corporations rather than with the government or universities. While some part of such files undoubtedly is truly proprietary, being privately generated and held for possible business advantage and hence not accessible to those outside the host institution, a large fraction is simply data already culled from a variety of sources, evaluated and placed in an organized format. No means exists for others to be aware of the existence of these files even when no proprietary element is present. Companies are also reluctant to make their data files publicly accessible for other reasons. There are costs involved in standardizing and cleaning up data prior to release as well as in the provisions of access itself. Finally, there is the intangible cost of possible liability associated with misuse or misinterpretation of the company's data by others.

Unpublished Character

Another factor preventing access to technical data is that, due to the pressure on space in journals, editors increasingly prevail upon authors to include only key or representative data. It also appears paradoxically that, while more data than ever before is being collected in a given experiment because of the efficiencies of computerized data logging, a smaller fraction finds its way into the literature because of the limited capability to interpret all observations and because of the aforementioned space constraints imposed by journal editors. A related difficulty is that frequently only derived or interpreted data will be published and all raw data (which might be more properly assessable or analyzed differently later by others) remain unpublished and inaccessible.

Military Classification

Military classification is a barrier to data access that most reasonable people can appreciate and accept. The problem as the writer sees it lies not so much with that body of data which at the moment has great military sensitivity but rather with the much larger body for which military classification is, or should be, no longer a barrier but which fails to be placed in an accessible place of record. The reasons are not hard to find: at the time of declassification, the original generators of the data are then scattered or diverted to other activities and funds which might have supported open publication of the work are no longer available.

Need To Know

"Need to know" can constitute a de facto barrier to data access even when no legal or military barrier exists. While in the social or business fields such rules or practices have their place, their application in the sphere of technology can only be seen as a product of excessive bureaucracy or of misguided attempts to reduce "nuisance" inquiries.

Transient Availability

Perhaps the most significant barrier is that designated as "transient availability" and created, ironically, by the power of new information gathering technologies. For example, a single satellite with its multi-channel sensors can generate as much as 2×10^{11} bits of data per day or again, literally miles of records are being accumulated annually from the world's exploratory drilling programs. Unfortunately, despite the recent prodigious gains in memory size, data compression, and storage cost reduction, there is no way all such data can be permanently archived and made accessible. Periodically engineers and scientists in charge of such files must make agonizing decisions as to what data to keep and what to discard. They do so, knowing full well that future needs, future interpretational abilities, or simply availability of time and space will likely later make that decision seem callous and irresponsible.

EXTRACTION OF DATA FROM ITS SOURCE

Data are recorded primarily on paper, to a lesser extent photographically and increasingly directly in some electronic or machine-readable form. Beginning with the paper medium, the first step is acquisition of the document via conventional library sources, reprints from the author, the Original Article Tear Sheet Service (OATS) offered by the Institute of Scientific Information, or electronic ordering of documents via the regular on-line bibliographic services. The enormous volume of modern literature, even in a narrowly defined field, has led to increased attention to automated methods of full text searching.(50) The American Chemical Society (ACS) has just announced that beginning this year, in collaboration with Bibliographic Research Services (BRS), the

full text of 18 of its research journals will be available on-line for search, retrieval and print. The file will initially contain 25,000 articles dating from 1980 to the present and will be updated every two weeks. Today these full text techniques require that the text and included data be digitized. Such a digitized record may already be available if the text at some point in the publication cycle has become the product of a word-processor or photo-typesetter. If not, the material must first be re-keyboarded or processed with an Optical Character Reader (OCR). Font-independent OCR machines are now available which are increasingly competitive economically. Once in machine-readable (digitized) form the document may be subjected to full text search with any of a number of commercially available search programs which in general include full Boolean matching, proximity search, truncation, fractional list matching and similar capabilities.

Problems, of course, with full-text searching are the large memory required and the relative slowness of search. A partial solution to the latter is provided by a system with a parallel array architecture of socalled associated processors, a group of large scale integrated circuits which perform simultaneous information scanning and matching. This approach affords search speeds in excess of 2×10^6 characters per second, more than 100 times the search speed of typical software sequential search technology. Since indices, keywords and directories are eliminated, substantial cost savings are realized and the shortcomings of imperfect intellectual assessment obviated. Queries are fed to an array of processors which simultaneously and independently compare the query with the stored text by character string matching with a string pattern matching algorithm implemented in each processor. One such commercially available system is General Electric's GESCAN2. Unfortunately, the associated processor technology has a major limitation with reference to searching of numeric data. Searching is restricted to Boolean "and/or/not" matching, and relational searches comparing numerical values to pre-stated limits are not feasible without drastic loss of search speed.

In many cases, for reasons of compact storage, electronic recording, machine searching, or convenient duplication of large volumes of data, the data are recorded and stored as magnetic tape or discs. These may be duplicated for dissemination or transmitted by any electronic communication link. Even though different codes, protocols or languages may have been used to create the magnetic record, these are usually now "translatable" with standard software packages.

Much technical data is captured in image form: photographically (e.g. astronomical photographs, medical radiology, or particle physics records) or graphically as a direct instrumental record of a parametric dependence of a quantity on time, voltage, spatial distribution, wavelength etc. Additionally graphics are frequently employed in a secondary treatment of data for visualization and comparison (e.g. phase diagrams or Pourbaix diagrams of electrochemical equilibria). The collection and cataloging of such data is a major task and involves both considerable intellectual as well as hand labor.(51) Development of techniques for searching this type of data storage is only just beginning. Photo images may be indexed, stored and retrieved by qualitative descriptors using a carefully structured facet analysis in conjunction with a controlled vocabulary.(52) Even when completed, such collections have short-comings. They are not readily updated and they require much specialist skill in their interpretation. For these and other reasons much effort has been expended recently to convert such photographic or graphic records to computer graphic form so that they might be more readily stored, searched, updated and displayed. Murray and Orser in their discussion of computer graphics as applied to phase diagrams(53) emphasize that computer graphics is not synonymous with digitization. Two distinct tasks must be performed: a) composition of the graph - the "drafting," and b) generation of the graphical representation from the available data and, in this case, the application of thermodynamic (topological) rules. Thus key feature of the graphs (lines, vertices, areas etc.) are not just implicit in a digitized image but through a relational database structure can be explicitly defined and manipulated by the computer in an efficient and natural way. An example of the result of the NBS phase diagram computer graphics program is shown in Fig. 4. The computer has not only "drafted" the graphics representation, it has also stored the original data shown and calculated the curves according to built-in algorithmic rules. Thus with acquisition of new data, the graphic representation of the whole body of data is readily updated.

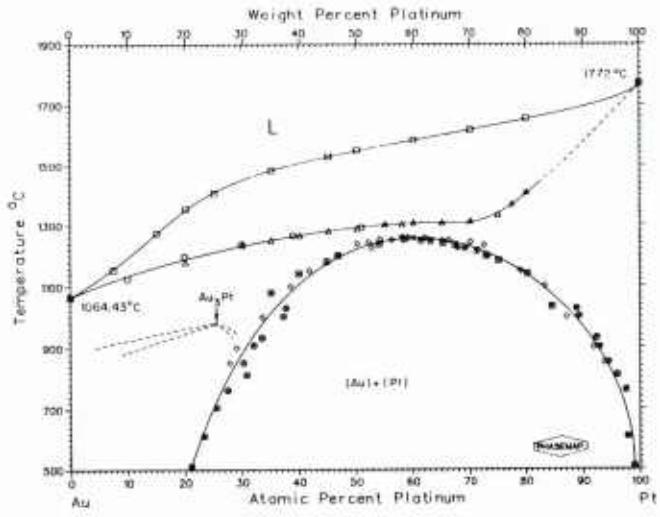


Fig. 4. Example of a computer drawn phase diagram wherein the computer has also been used to calculate the shape and position of lines representing the experimental data [the Au-Pt system after Singhal, S.P. Bull. Alloy Phase Diagrams 2 (1) (1981) 66; see also Murray and Orser (53)].

The technology in the image storage field is still in a considerable state of flux.(54) One of the most promising newer techniques is that of videodisc technology which has capability for both digital and image storage.(55) Present video discs have a storage density 1000X greater than magnetic media at a fraction of the cost. Commercial products offer a storage of 2 x 10^{10} bits per side i.e., several 1000's of pages of information. Unfortunately the most popular display unit, the CRT, falls short by about a factor of 4 in supplying the necessary resolution to read text. Schipma and Becker(56) are exploring one way out of this dilemma. They divide the printed page of text into 4 to 6 pieces depending on format and store each of these pieces on one frame. The feasibility of a special device they have developed to fragment each page by computer algorithm has been demonstrated. Toshiba and GE, among others, have been working on a more direct hardware solution: the development of a CRT display capable of 2048 x 2048 pixels per frame, more than 16X the capability of the commercial TV screen.

Another limitation of the video disc is the fact that at present it cannot be erased, i.e., it is a "write once" or "read only" device. Shamir and Rivest(57) have proposed a novel but simple solution based on the very high information storage density of the video disc. That is, they propose to sacrifice a portion of the total available space and spread these unused spaces throughout the disc. Thus it is possible to enter updating information in the blank spaces without formally erasing the old obsolete data.

KNOWLEDGE ORGANIZATION

If a collection of numerical and factual data is to be more than a random compilation of numbers and words - in short, if it is to become an information system, these data must be subjected to some sort of intellectual structuring. Broadly speaking there are three models for organizing information:

the hierarchical model describing one-to-many relationships between entities

the network model describing many-to-many relationships between entities

the relational model which represents relations in the form of a two-dimensional table

Combinations of all these are of course possible. Which model to adopt in designing an information system depends on both the intrinsic character of the body of knowledge to be organized and in part on the use or uses to which it is to be put. Dubois(58) has described numeric database design in other terms, namely the representation of the interrelationship between patterns (i.e. chemical compounds or other entities) and their properties. He asserts that any effective representation must exhibit two characteristics:

it must be generative: the patterns and their properties must be retrievable from the database starting from the descriptors

it must be generic: the descriptors must give rise to an organization of the patterns and of their properties into the database

A particular application of this concept to the field of chemistry is achieved by means of valued chromatic graphs wherein the graph represents the chemical topology and the chromatism the atoms and bonds.

If the information system is computerized, still other features of any organizational scheme are forced by the computer system. Technical information systems typically handle very large volumes of data but perform only a few and simple calculations on the data stored. Since most of the data must be stored in backup (tapes, discs, drums, etc.) and only small portions brought to the main memory of the computer for processing at any one time, the effective speed of the machine is constrained by the required frequency of access to backup memory and the low rate of transfer from backup memory to main memory rather than by the computational speed of the CPU. Thus, the objective in computerized file structure design is to arrange the access paths such as to reduce number of "look-ups" and the amount of data that must be transferred for each task. Once the logic of the organization of the data collection is fixed, following any of the principles outlined above, then it is possible to use the computer itself to derive an unambiguous sort key which can then be used to sort the access elements of the data file in such a way that the most closely related elements are stored nearest to each other. Yang(59) has reviewed many of the methods used in computer file structuring including tree structures, hashing, division into sub-files, cluster trees, and arrangements for near-neighbor searching.

It should perhaps be remarked that in all of the computer organizational schemes discussed thus far, much of the nature of the file structure is inferrable by the user from the various menus, keyword lists and prompts which are presented to the user. At a somewhat higher level of sophistication in system design, the user directs a virtually free form, natural language question to the system which is then transformed, via automatic invocation of thesauri and syntactical analysis, to a form matching the hidden structure of the system. We are just beginning to see the first embodiments of this type of information system in the numeric data field. The ultimate, of course, will be the realization of socalled "expert" systems(60)(61) which not only admit very free forms of questions but will automatically prompt and interrogate the user so as to progressively narrow the range of questions and answers by inference and deduction. In a sense such systems will be "self-learning" in that with use they will refine and elaborate their organization so that future answers will be delivered both more quickly and less ambiguously.

The previous discussion of data extraction dealt only with the mechanics of digitization of information from whatever form it appeared in the original source. Now we wish to consider the possible choices for storage of the information. As an example suppose we are confronted with a set of data for the emf of a thermocouple as a function of temperature. This might appear graphically as shown in Fig. 5. Experience has also revealed that such data sets may frequently be represented by an expression such as

$$\varepsilon = A + BT + CT^2$$

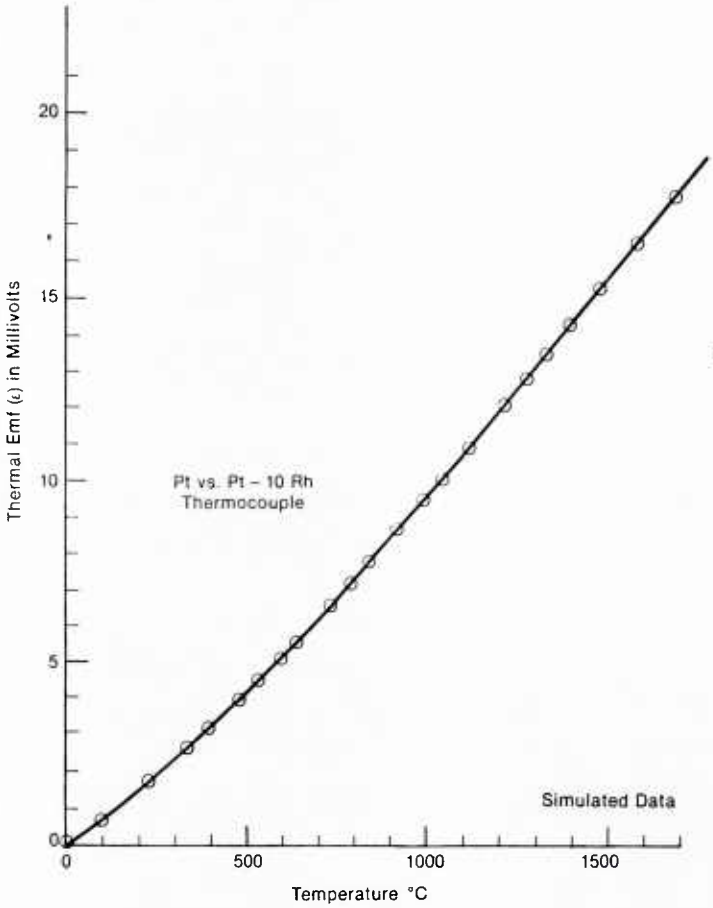


Fig. 5. Simulated set of experimental data (thermal emf vs. temperature) to be computer stored for later retrieval and display.

In considering what is to be stored in the computer the choices are:

a) store a set of raw data points representing the results of an experimental study

b) store pairs of values of ε and T at regular intervals as read from a smoothed curve

c) store values of A, B and C in the parametric expression (empirical or theoretically based), the expression itself, and the algorithm for performing the calculation

d) store coordinate values of an arbitrary number of points arbitrarily spaced along a smooth curve representation chosen such as to reproduce the curve with desired precision

Choice a) will be favored by the data analyst and researcher but will increasingly be impractical for storage in the main memory as additional data sets accumulate for the same property of the same material. Choices b), c) and d) will all permit graphic display of the data set with whatever degree of resolution is desired. Choice b) may have some utility for direct retrieval of values at frequently needed intervals of the independent variable (say 100° steps of temperature in the example). With choice b) and d) the calculational power of the computer may be invoked to interpolate values of the dependent variable at any chosen value of the independent variable. The reliability of the interpolation, however, obviously rests on whether a linear functional relation is assumed between points, some established parametric function as in c) or that implied by an ad hoc curve fitting procedure as in d). Each option represents a different trade-off between data compression and retrieval speed. The information system must be designed such that it is very clear to the user which choice of data storage has been made in each case.

COMPILATION OF DATA

Prior to the evaluation and application of numerical or factual data, the compilations resulting from the preceding processes outlined above must undergo certain other

manipulations. We include here by way of example: homogenization, data reduction, and smoothing of data.

Homogenization - Data as reported in the original literature, even for the same property of the same material, will in general be stated in different units; be measured from different tests and under somewhat varying conditions of common independent variables such as temperature, pressure and the like; and perhaps be derived from raw data using different calculational methods and different fundamental constants. If different data sets from different investigations are to be intercompared, a "best" or recommended set of values chosen for inclusion in a reference database, and reliability established, numerous adjustments have to be made. Such matters have been discussed by Lide and Rossmassler,(62) Lide and Paul(63) and Ho and Touloukian.(64)

Data Reduction - It may be determined that a different type of analysis of the raw data from that originally reported is desirable before entry to a reference database or, very commonly, that memory space conservation or ease of application of the results demand that only certain selected values be stored. For example, in measurements of the mechanical phenomenon called "creep" in metals the experimenter measures deformation as a function of time under load at constant temperature. However, depending on the intended application of the data, the database manager may choose to store, not the full strain-time relationship, but only one or several of the following parameters:

secondary creep rate	$d\varepsilon_2/dt$
strain for secondary creep	ε_2
time for secondary creep	t_2
tertiary creep rate	$d\varepsilon_3/dt$
strain for tertiary creep	ε_3
time for tertiary creep	t_3
strain at failure, etc.	t_f

All of these desired parameters are defined in terms of slopes, coordinates, and inflection points for various parts of the total strain-time curve, as shown schematically in Fig. 6.

Smoothing of Data - Preliminary to detailed evaluation of data are such steps as the discard of outliers, regression analysis and statistical analysis. These matters are discussed in detail in standard texts and monographs on applied statistics. A brief summary and useful reference list have been provided by James.(65) Especially to be noted are the numerous computer programs that have been designed specifically for data analysis, smoothing and interpolation.(66-70)

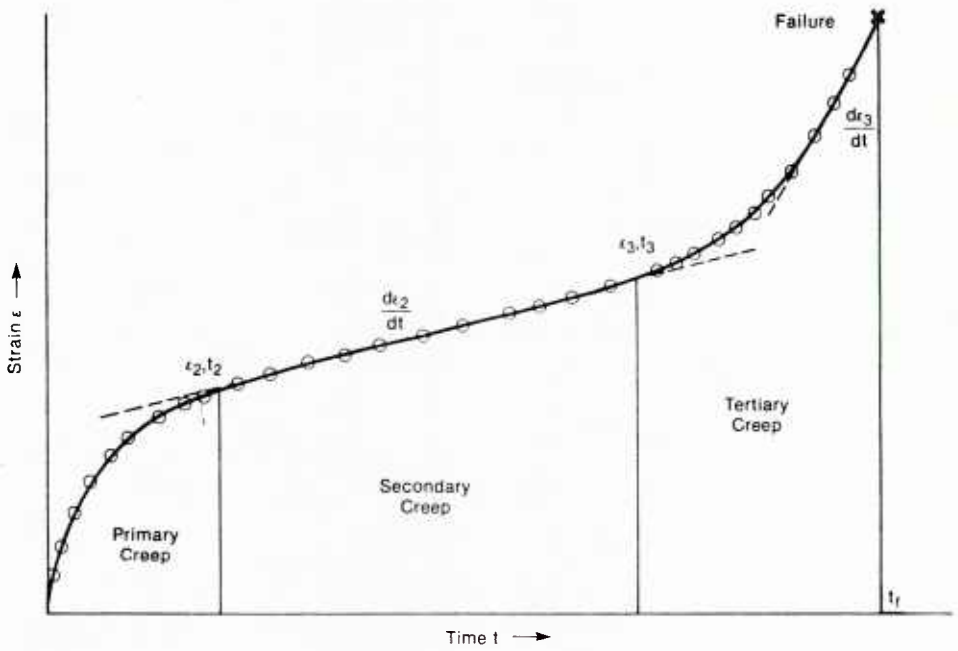


Fig. 6. Simulated creep curve (elongation vs. time) for high temperature deformation of metal under constant stress and temperature to illustrate distinction between raw and derived data.

Standardization - Problems in accomplishing the various tasks just outlined will be progressively eased by the increasing standardization of test methods, of reference materials, of data reporting techniques and of machine-readable numeric records. Test methods are standardized by international groups such as the International Standards Organization or the International Electrotechnical Commission and nationally, for example, by such organizations as the American National Standards Institute in the U.S. and the Deutsches Institut für Normung in the Federal Republic of Germany. Standard reference materials and standard reference data are two other important developments in recent years. Availability of these artifactual and numerical standards facilitates calibration of test equipment, intercomparison of laboratory data, identification of materials and structures and analysis of other types of data. These topics have been discussed by Westbrook(71) and by Berger and Tucker.(72) A major difficulty in analyzing and using data from the original literature has been the generally inadequate and inconsistent reporting of test materials, test methods, measurement conditions, data reduction techniques, and error estimation. This topic has been discussed in a general way by Rossmassler.(73) Valuable contributions to alleviation of the problem have been made by ASTM(74) and by CODATA in a series of publications.(75)

A final aspect of standardization pertinent to numeric data lies in the field of communication. Data exchange, amalgamation of files and inter-system connection have in the past been seriously impeded by incompatability of the formats and protocols used in digital recording of numeric data. A proposal(76) now under study by the American National Standards Institute, derived from an ISO bibliographic standard, may overcome this difficulty. Early test applications of this tentative standard have been reported(77)(78); Staiger and Wheeler(79) have suggested a standard for print publication of tabular displays of numeric data which should facilitate their adoption into machine-readable form.

CONCLUSION

The processes of extraction and compilation of numerical and factual data have acquired an unprecedented significance and importance in modern society. Two general observations may be made with reference to this development. The first is the pervasive and powerful influence of the computer. While it eases many mechanistic problems, it also imposes on its users new strictures of logical development and formal structuring of information. Secondly the enormity, complexity and urgencies of the job to be done are enforcing new modes and higher levels of cooperation: between data generators and data users, between the public and the private sectors, between technical disciplines, and between the nations of the world.

REFERENCES

1. Price, D. deS, "Science Since Babylon," Yale Univ. Press, New Haven (1961).

2. Williams, M.E., 1980, "Database and Online Statistics for 1979," Bulletin of the American Society for Information Science, 1980 December; 7(2): 27-29.

3. Bartkus, E.P., private communication.

4. Buck, E., "The Industrial Data Bank: Utopia vs the Real World" fm Phase Equilibria and Fluid Properties in the Chemical Industry, ACS, p 455-458.

5. Westbrook, J.H., "EMPIS: A Materials Data Program of an Electrical Manufacturing Company," Data for Science and Technology, Proceedings of 7th International CODATA Conference, Oct. 1980, Kyoto, Japan, Pergamon Press (1981) 462.

6. "Index to Scientific Reviews" Institute for Scientific Information, Philadelphia (1979) 7.

7. Chen, C.C., "Scientific and Technical Information Sources," MIT Press, Cambridge, MA (1977).

8. Kharasch, N., Wolf, W. and Harrison, E.C.P., eds., "Index to Reviews, Symposia Volumes, and Monographs in Organic Chemistry," Pergamon, New York, 1962, 345 pp. Supplements 1964 260 pp and 1966 326 pp.

9. National Advisory Committee on Research in the Geological Sciences (Canada), "A National System for Storage and Retrieval of Geological Data in Canada," Ottawa, Canada, Geological Survey of Canada, 1967, p 23-24.

10. Lerner, R.G., "Access to Data in the Primary Literature" in Data Handling for Science and Technology, S.A. Rossmassler and D.G. Watson, eds., North Holland, Amsterdam (1980), p. 75.

11. Murdock, J.W., "Current Knowledge on Numerical Data Indexing and Possible Future Developments," NSF Report April 1978.

12. CODATA Bull. No. 12 (Sept. 1974), Energy Data Accessing and/or Retrieval (Report on Data Tagging, compiled by a Panel of Experts at the Energy R & D Data Workshop held in Gaithersburg, MD, USA, May 1974), 12 pp.

13. CODATA Bull. No. 19 (June 1976), Flagging and Tagging Data (Report of the ICSU AB/CODATA Joint Working Group), 22 pp.

14. Hawkins, D.T., "Problems in Physical Data Retrieval," J. Chem. Info. Computer Sci. 20 (1980) p 143-145.

15. Wood, B.L., "Review of Scientific and Technical Numeric Data Bases," report by King Research, Inc., Rockville, MD (1977).

16. "Data Activities in Britain," Dept. of Education and Science, 3rd ed, London 1969.

17. Williams, C.H., "Guide to European Sources of Technical Information," 3rd ed, Guernsey 1970.

18. "4th Consolidated Guide to International Data Exchange Through the World Data Centers" ICSU, Washington (1979) 113 pp.

19. International Compendium of Numerical Data Projects, Springer-Verlag, Heidelberg-Berlin (1969) 295 pp.

20. Emptoz, G., ed, Inventory of Data Sources in Science and Technology: A Preliminary Survey, Unesco Press and CODATA Secretariat, Paris (1982).

21. Westbrook, J.H. and Desai, J.D., "Data Sources for Materials Scientists and Engineers," Ann. Reviews of Materials Science 8 (1978) 359.

22. Schönberg, M., "Data Gaps in Respect to Organic Industrial Chemicals," in Data for Science and Technology, Proc. 7th Int'l. CODATA Conf., P. Glaeser, ed, Pergamon (1981) 447.

23. Hampel, V.E., Henry, E.A., Kuhn, R.W. and Lyles, L., "Acquisition, Storage, Retrieval, Utilization and Display of Computerized Data in LLL Data Bank of Physical and Chemical Properties," in Proc. 5th Int'l. CODATA Conf. Pergamon Press (1977).

24. Merrill, D. and Austin, D.M., eds, "ERDA Interlaboratory Working Group for Data Exchange," Annual Rpt for 1976. Lawrence Berkeley Lab, U. Calif., 123 pp.

25. Loebl, A.S., et al, "Regional Information Group (RIG) Energy Environmental and Socioeconomic Data Bases and Associated Software at Oak Ridge National Laboratory," ORNL/TM5600 (1976) 125 pp.

26. Edwards, M.D., "States of the National Water Data Exchange (NAWDEX)," Dept. of Interior, Geological Survey (1976) 23 pp.

27. Noe, C.D., "ENDEX - A System for Locating Historical Environmental Data," Proc 5th Int'l. CODATA Conf., Pergamon (1977).

28. Oak Ridge National Laboratory, "National Inventory of Biological Monitoring Programs," Summary Report, Oct (1976) 538 pp.

29. Carroll, B.T. and Maskewitz, B.F., "Information Analysis Centers," Annual Review of Information Science and Technology 15 (1980) 147.

30 Williams, M.E., et al, "Computer-Readable Data Bases: A Directory and Data Sourcebook." Washington, DC: American Society for Information Science; 1982. 1500 pp Available from: Knowledge Industry Publications, Inc., White Plains, NY.

31. Landau, R.N., Abels, D.M., Wanger, J., "Directory of Online Databases," Santa Monica, CA, Cuadra Associates, Inc., 1523 Sixth St., Suite 12, Santa Monica, CA 90401; v.4, #3, Spring (1983).

32. Wanger, J., Landau, R.N., 1980,"Nonbibliographic Online Database Services", Journal of the American Society for Information Science, 1980 May; 31(3): 171-180.

33. Luedke, J.A., Jr., Kovacs, G.J., Fried, J.B., 1977, "Numeric Data Bases and Systems." In: Williams, M.E., ed. Annual Review of Information Science and Technology: Vol. 12. White Plains, NY: Knowledge Industry Publications, Inc.; 1977, 119-181.

34. Luedke, J.A., Jr., "Numeric Databases Online," On-Line Review, Sept. 1977 1(3).

35. Fried, J.B., "On-Line Numeric Databases," Bull. Amer. Soc. Info. Sci., Feb. (1975) 17.

36. Tomberg, A., "European Information Networks," Annual Review of Information Science and Technology," 12 (1977) 183-216.

37. Heller, S.R. and Milne, G.W.A., "The NIH-EPA Chemical Information System," in Data for Science and Technology, Proc. 7th Int'l. CODATA Conf., P.S. Glaeser, ed. (1981) 343.

38. Westbrook, J.H. and Rumble, J.R., "Computerized Materials Data Systems," Proceedings of the Materials Data Workshop, Fairfield Glade, TN (1982) 133 pp.

39. Hampel, V.E. et al. "An Intelligent Gateway Computer for Information and Modeling Networks" Lawrence Livermore National Laboratory, Preprint UCRL-5439, Aug. 1983.

40. Williams, M.E. and Preece, S.E., "A Mini-Transparent System Using an Alpha Microprocessor," Proc. 2nd On-Line Meeting (1981) 499.

41. CODATA Bull. No. 31, "Data Needs for Energy," (Mar. 1979), 30 pp.

42. Weisman, H.M., "Needs of American Chemical Society Members for Property Data," J Chem Doc 7 (1967) 9-14.

43. Stockmayer, W.H., et al, "National Needs for Critically Evaluated Physical and Chemical Data" (CODAN Report) Nat. Acad. of Sciences, Washington (1978).

44. Transnational Data Report, March 1980 2(8) originally in "AFIPS Report on Transborder Data Flows."

45. Schiller, H.I., "Who Knows: Information in the Age of the Fortune 500," Ablex Publ. Corp., Norwood, NJ (1981).

46. Corson, D.R., et al, "Scientific Communication and National Security," NAS Report Sept. 1982.

47. Keplinger, M.S., "Copyright and Information Technology," Annual Review of Information Science and Technology 15 (1980) 3.

48. Fivozinsky, S., working paper prepared for the CODATA ADD Task Group, 1983.

49. McDonald, D.D., Rodger, E.J. and Squires, J.L., "Findings of the IFLA International Study on the Copyright of Bibliographical Records in Machine-Readable Form," submitted to IFLA June 1982 by King Research, Inc.

50. O'Connor, J., "Data Retrieval by Text Searching," J. Chem. Inform. Computer Sci. 17.3 (1977) 181.

51. Lunin, L.F., et al, "Organizing for Information Interaction in a Radiology Department: Focus on Image Analysis, Storage, and Retrieval," Proc. ASIS Annual Meeting 19 (1982) 179.

52. Batty, D. and Stevens, P., "Automated Retrieval Systems for Photo-Image Collections: Problems and a Solution," Proc. ASIS Annual Meeting 19 (1982) 23.

53. Murray, J.L. and Orser, D.J., "Interactive Computer Graphics for Storing Phase Diagrams," Bull. Alloy Phase Diagrams 1 (1) 19.

54. Turtle, H., Penniman, W.D., and Hickey, T.B., "Data Entry Display Devices for Interactive Information Retrieval," Annual Rev. Info. Science and Tech. 16 (1981).

55. Marsh, F.E., "Videodisc Technology" J. Amer. Soc. Info. Sci., July 1982, p.237.

56. Schipma, P.B. and Becker, D.S.,"Text Storage and Display via Videodisc or 'Someday my Prints Will Come' ", Proc. 2nd Outline Meeting, (1981) 427.

57. Shamir, A. and Rivest R., reported in Ind. Research and Development, Feb. 1983, p.70.

58. Dubois, J.E., "Computer Representation of Numerical and Graphic Data," Proc. 3rd Int'l. CODATA Conf., Le Creusot, France (1972) 58.

59. Yang, C.S., "Design and Evaluation of File Structures," Annual Rev. Infor. Science and Tech., 13(1978) 125.

60. Smith, L.C., "Artificial Intelligence Applications in Information Systems," Annual Rev. Info. Science and Tech., 15 (1980) 67.

61. Duda, R.O. and Shortliffe, E.H., "Expert Systems Research," Science, 220 (1983) 261.

62. Lide, D.R. and Rossmassler, S.A. "Status Report on Critical Compilations of Physical Chemical Data," Annual Reviews in Physical Chemistry, 24 (1973) 135.

63. Lide, D.R. and Paul, M.A. eds., Critical Evaluation of Chemical and Physical Structural Information, National Academy of Sciences, Washington, D.C. (1974) 628 pp.

64. Ho, C.Y. and Touloukian, Y.S. "Methodology in the Generation of Critically Evaluated, Analyzed and Synthesized Thermal, Electrical and Optical Properties Data for Solid Materials," Proc. 5th Int'l. CODATA Conf., Boulder, CO (1976) 615.

65. James, G.D., "Analysis and Interpretation of Data" in Data Handling for Science and Technology, Rossmassler, S.A. and Watson, D.G. eds., North Holland, Amsterdam (1980) 55.

66. Daniel, C., Wood, F.S. (1971) Fitting Equations to Data: Computer Analysis of Multifactor Data for Scientists and Engineers, New York: Wiley 342 pp.

67. McNeil, D.R., (1976) Interactive Data Analysis, New York: Wiley 181 pp.

68. Hahn, G.J., Nelson, W.B., Celtay, C., "STATSYSTEM-A User-Oriented Interactive System for Statistical Data Analysis," Proc. Am. Stat. Assoc. Stat. Comp. Sect., (1975) 118.

69. Fried, J.B., "BASIS: On-Line Retrieval and Analysis of Large Numeric Data Bases," Proc. 5th Biennial Int. CODATA Conf., Boulder Colorado, 1976, pp. 541-47.

70. Library Software Index and Network Software Services Program Index, May 1975, Rockville, MD: General Electric Co. 64 pp.

71. Westbrook, J.H., "Materials Standards and Specifications" in Kirk-Othmer Encycl. of Chem. Technology, 3rd ed., V 15, (1981) 32.

72. Berger, P.W. and Tucker, J.C., "Standards and Guidelines for Data" in Data Handling for Science and Technology, Rossmassler, S.A. and Watson, D.G., eds., North Holland, Amsterdam, (1980) 93.

73. Rossmassler, S.A., "Presentation of Data in the Primary Literature" in Data Handling for Science and Technology, Rossmassler, S.A. and Watson, D.G. eds., North Holland, Amsterdam, (1980) 65.

74. Manual on Presentation of Data and Control Chart Analysis, 1976. ASTM Special Technical Publication 15D. Philadelphia, PA: Am. Soc. Test. Mater. 186 pp.

75. "Guide for the Presentation in the Primary Literature of Numerical Data Derived from Experiments" CODATA Bull. #9, (1973) 6 pp. and since supplemented by specialist area guides:

Bulletin No.	
13	Chemical Kinetics Data (1974) 8 pp.
15	Biochemical Equilibrium Data (1975) 32 pp.
25	Biological Data (1977) 5 pp.
30	Phys. Property Correlations & Estimation Procedures (1978) 6 pp.
32	Observations in the Geosciences (1979) 6 pp.
44	Calorimetric Measurements on Cellular Systems (1981) 8 pp.
46	Astronomical Data (1982) 9 pp.
47	Thermodynamic Tables (1982) 13 pp.

76. American National Standards Institute (ANSI). Subcommittee X3L5 on Data Formats and Labeling, Draft Proposal: American National Standard Specifications for an Information Interchange Data Descriptive File, 1977 March 2. 62 p. (Working paper; X3L5/589F-2-28-77 (corrected)). Current version available from Mr. A.A. Brooks, P.O. Box X, Oak Ridge, TN 37830.

77. Benkovitz, C.M., McNeely, B.N. Wiley, Richard A., User's Guide for the IWGDE Level 1 Implementation of the Proposed American National Standard Specifications for an Information Interchange Data Descriptive File, Los Alamos, NM: Los Alamos Scientific Laboratory; 1977 March. 20 p. (Rough draft).

78. Merrill, D., Austin, D.M., eds., ERDA Interlaboratory Working Group for Data Exchange (IWGDE): Annual Report for Fiscal Year 1976. Berkeley, CA: University of California Lawrence Berkeley Laboratory, Technical Information Division; 1976 September 9, 123 p (LBL-5329). Available from: NTIS.

79. Staiger, D.L., Wheeler, W., "Capturing Numeric Data in Machine-Readable Form and Generating Graphic Output Displays," in Dreyfus, Bertrand, ed. Proceedings of the 5th International CODATA Conference on Generation, Compilation, Evaluation and Dissemination of Data for Science and Technology; 1976 June 28-July 1; Boulder, CO. Elmsford, NY Pergamon Press; 1977 10 p.

THE EVALUATION/VALIDATION PROCESS - PRACTICAL CONSIDERATIONS AND METHODOLOGY FOR THE EVALUATION OF PHENOMENOLOGICAL DATA

Dr Anthony J. Barrett,
Chairman,
ESDU International Ltd,
251-259 Regent Street,
LONDON, W1R 7AD.

SUMMARY

Scientists and engineers strive to ensure that their work is based upon objective principles and that it is repeatable to close tolerances. The factual and numerical data resources which are available to them, however, do not always assist this intention particularly where data are being used as a basis of decision in the engineering design process which is directed at the realisation of a practical product goal. Subjective influences, related to imparted or acquired skills and experience, often apply in such cases. These have to be taken into account during the refinement processes, (evaluation and validation), which need to be undertaken during the construction of numerical and factual data bases. The practical consequences of inadequately refined data are reflected in unnecessary costs and uncompetitive product performance. Careful management of data refinement is always needed and can be seen to be increasingly important as a greater proportion of data is stored electronically or becomes embedded in computer aided engineering and design systems.

INTRODUCTION

The world is involved in a technological revolution. It has been likened to that earlier revolution when the major industrial nations, as we know them today, became pre-occupied with the application of new energy sources together with new mechanical devices so as to transform the world's manufacturing and transport facilities. The more recent revolution relates to information technology and many of us are pre-occupied with the application of electronic devices to transform the storage, retrieval, communication and application of the world's information resources. A great deal of the effort of 'information technology' is being applied to the quality and development of the hardware technology; far less is being applied to the quality and development of the information itself. Yet the outcome of the 'information technology revolution' will be sterile if our ability to store and manipulate increasingly vast quantities of information is not matched by our ability to certify the quality and relevance of the information so handled. Simply processing, or having easier access to, more and more information does not mean that one is better informed - it may merely offer such a bewildering range of alternatives that effective decision and action are stultified.

Numerical and factual data bases, whether they are committed to printed paper or are embodied in an electronic system, constitute a particular class of information resource. They are of primary importance to industry and defence because they provide an immediate basis for reasoning and decision making. They are different, for example, to bibiliographic files which essentially provide direction to sources of specific information. These sources have to be retrieved and the contents read and absorbed before a basis for reasoning and decision can exist. The numerical or factual data base, by contrast, is often the last stop on the path towards making a decision. When it is consulted it should only reveal decision alternatives that are valid and for each should only provide relevant information that has substantiated value or is of known quality. This quality is established by the evaluation and validation processes applied during the construction of such a data base. We shall be looking at these processes in relation to a particular sub-set of numerical and factual data which is here described as "phenomenological data". An engineering application will generally be assumed although many of the principles examined will apply elsewhere.

First of all some explanation of terms is needed in order to clarify the scope of our discussion.

TERMINOLOGY

The terms 'data' and 'information' are increasingly used interchangeably. It is still advisable to refer to data in the strict sense of factual information used as a basis for reasoning or calculation. Engineering data thus will be taken to include the numerical values, such as physical constants, theoretical and empirical coefficients. However, such numerical values are usually inseparable from the design or analytical algorithms with which they are used and which are part of the same resource of knowledge.

Take, for example, the simple subsonic drag equation, for a body of reference area S in a fluid stream of velocity V:-

$$D = \tfrac{1}{2}\rho V^2 S C_D$$

The physical constant ρ (fluid density) has an existence quite apart from that in this formulation. The empirical coefficient C_D, by contrast, only has relevance in connection with this equation or some mathematical derivation which, in effect, infers the equation. The equation itself has to be modified in form as do the parameters it embraces when a different speed regime is encountered. Such constants, coefficients, and the equations themselves are of vital importance in the decision making process of design and all of them constitute 'data' within the terms used in this paper.

A physical constant, such as air density, will generally have an accepted method for its determination. That method will be well documented, be reproducible to within close tolerances and will be little influenced by subjective considerations. A parameter such as C_D, however, may not be determined by well standardised methods. The apparatus used in its determination from a wind-tunnel test, for example, is not standardised, it will not replicate exactly the conditions of flight, all manner of corrections may be needed and the very circumstances of the test and its interpretation may depend heavily upon the experience and skill of those involved in running wind-tunnel experiments. In engineering data there is clearly much scope for subjective influence.

Data, using the term in the broad sense we are using it here, which are affected by expert knowledge and skill based upon experience and training, require special consideration in any process which is aimed at evaluating them and validating their use in engineering practice. Such data have been termed phenomenological; the means of evaluating and validating them have some degree of subjective influence. Whilst for most engineering data there is a theoretical basis to some level of refinement, this level is by no means uniform across all the data that are needed in the engineering design process.

This is a suitable point at which to look briefly at the design process. The circumstances of the intended application of data are a factor to be taken into account in their evaluation and validation.

APPLICATION OF DATA IN DESIGN PROCESSES

It is a sound precept of modern business life to start from the viewpoint of the customer. The customer, or user, of the data bases we construct is engaged in the application of data to some purpose. Data bases provide the foundation of decision at some point in a design process. Alternatively they provide the foundation on which judgements of an existing design may be based either to forecast whether that design meets some predetermined criterion of performance or to explain why the design has failed in service to perform in the way expected.

In making decisions the engineer has many things to take into account. The consumer for the product being designed defines a specification which lays down the required performance of the product. On this specification the manufacturer will superimpose cost targets, the company philosophy (somewhere in the spectrum which runs from high quality/high price tag to planned obsolescence) and the manufacturing/market time scale. Also upon this specification there are increasingly being overlaid legislative requirements concerning safety, environmental acceptability and energy saving. The designer must work within all these boundaries.

In the real world the specification, in the broad sense in which I am using the term, inevitably changes during design and manufacture but we shall ignore this in the interests of simplifying our study. From the basic specification there is no unique process of design to be followed. However, there are two broad classes of activity into which, or between which, most design activities fall and have been described in some detail in references 1 and 2. One of these activities I call "development design" and the other "critical design".

In development design a more or less successful existing product is scaled up or tailored to new requirements on an almost empirical basis. This process draws heavily upon the engineer's mechanical sense and previous practical experience. The product preceding the one to be designed is its father and mother, laboratory and test house. The physical laws governing the performance of the new product are modelled in the product which it supersedes to a degree which is quite adequate if the customer does not call for nonlinearly scaled performance increases (leading to over-development), if competitors do not start to offer novel features in their products, if materials costs do not fluctuate dramatically, if the same energy resources remain continuously available and if the scaling up of any environmental damage which the product can cause remains acceptable to society.

In an attempt to remove some of the uncertainties of design purely by the development process, engineers have recourse to the quite different process of critical design. This is illustrated in much simplified form in Fig. 1. The process draws on the engineer's mechanical sense and practical experience as well as on similar products of which the engineer is aware. It requires intuitive and creative flair to produce, first of all, a more or less novel concept in response to the original specification. This concept will be set down first in schematic form. Then the engineer, or teams of engineers, will undertake an essentially intellectual exercise. In this the schematic is analysed quantitatively against as many of the physical laws with which it is known the finished product must comply. Engineers aided by whatever computer power is available, simulate and test every aspect of performance which the product is to deliver, simulate and test the effect of the product on the environment and so forth. Having found where the original schematic is inadequate, for example in terms of performance, materials usage, or cost, the schematic is refined. Then the process of analysis is repeated until after perhaps many iterations there is sufficient confidence for a prototype to be built and tested under more or less representative service conditions.

There are other processes of design. Ad initio design or true synthesis goes a stage beyond critical design by removing the need for intellectual intervention at all stages beyond the specification (or some part of the specification). The computer is now frequently employed to go around the same iterative loops in critical design as those previously followed by human hand. But I do not regard this as true synthesis. Such processes are as yet rare though examples exist, such as those originally propounded by Mitchell, Cox & Hemp for structural design (3). Synthesis is also approached in the 'inverse design' process for aerofoils (4); here the specification of an upper surface pressure distribution enables exchange rates between parameters including drag rise Mach number, lift coefficient and aerofoil thickness to be forecast. Although developments such as these may one day transform the design process, for some time yet to come the iterative loops in the critical design process will remain as an essential feature of the engineer's work.

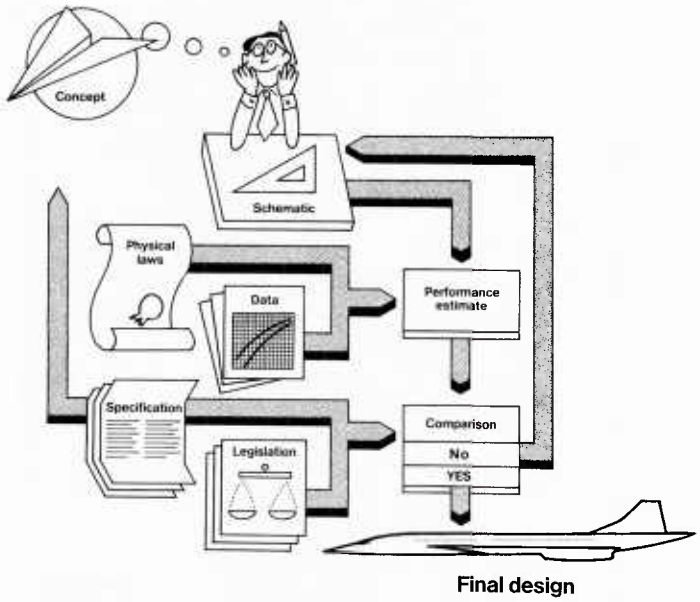


FIGURE 1 CARTOON OF CRITICAL DESIGN PROCESS

Returning to our 'cartoon' of the design process, illustrated in Fig. 1, we see that many types of information are embedded in it. When the process was, and in many cases still is, carried out mainly by human effort there were complex interactions between these information 'banks'. During the process of iterating around the loops the data available would come before intelligent inspection and, albeit intuitively, the human operator would exercise some degree of judgement on the relevance and quality of the data in relation to the application. Inconsistencies would be noted, judgements on the degree of quality needed would be made in terms of the effect which, for example, low accuracy data might have on the integrity of the design although these may not always be apparent.

Some years ago I was involved with a group of engineers debating the process of estimation of aircraft drag. At one point I was prompted to enquire what accuracy was

needed in the prediction of aircraft drag since the answer to that question would have an important influence on the cost and time involved in making design predictions. After a little discussion the designers in the group were agreed that an accuracy in the region of 1½% was needed; at that time, competitive aircraft were being marketed on the basis that a 1½% difference in performance could make all the difference in obtaining a contract or losing it. Among the group were also a number of wind-tunnel experts who immediately made the response that tests for many of the wind-tunnel data which would be used in drag prediction were not even repeatable to better than 4%!

The exposure of inconsistencies of this type was more likely when the detail of the design process was more directly under human expert surveillance than can possibly happen at the present stage of development of computer aided engineering and design. We shall consider the consequences of computer design methods again later. Suffice for the present to note that the key tools in design are a sound knowledge of the physical laws affecting the performance of a product and of the numerical data which enable those laws to be applied. Clearly these tools must be 'sharp' when applied by human hand and they must be even 'sharper' before they are to be embedded in computer aided design systems.

JUDGEMENTS APPLIED TO DATA AT THE TIME OF APPLICATION

In the experience related above, concerning the discussion on drag estimation, it was clear that the designers should have been making judgements on the quality of the data derived from wind-tunnel testing in relation to their application and its effect on the validity of the decisions they were making. To redress the balance in the designers' favour let us look briefly at the sort of judgement they are frequently more able to make and the consequences.

By way of example we consider the simple type of expression, frequently used in design work, to evaluate the elastic, uniaxial buckling stress of a thin sheet panel.

$$f_b = K \frac{E}{(1-\sigma^2)} \left(\frac{t}{b}\right)^2$$

In this the buckling stress, f_b, is related to buckling coefficient K, elastic modulus E, Poisson's ratio σ and the geometrical values of plate thickness and width, t and b respectively. There are two items of materials data, E and σ, each of which is derived by well established experimental methods. In the case of E, values should be provided in such publications as (5) and (6) to carefully prescribed statistical tolerances. Poisson's ratio, by contrast, is often available only to more crudely defined tolerances and at worst is often taken as 0.3 for aluminium alloys or 0.25 for steel! This is not as inconsistent as may at first appear for, whilst a 10% variation in E would lead to a 10% error in the estimated buckling stress, a similar variation in σ has but a 2% effect. Here the designer, and the data provider, could make judgements on the quality of data needed which are compatible with the known physical laws on which the relationship is based; the judgements are not subjective.

By contrast the buckling coefficient K may well be affected by subjective judgements particularly if it is based heavily upon experimental results. K is a function of many other parameters - it may almost become a depository for all other factors, known and unknown. In addition to plate aspect ratio the value of K is affected by edge restraint conditions, out-of-flatness of the plate and internal stress. This leaves plenty of scope for the application of judgement and experience in cases which may not be open to theoretical evaluation or where practical details of plate manufacture and construction differ from the idealisations used in experimental or theoretical evaluations.

Every one of the circumstances of the actual application of the data in design and analysis has its mirror image in the judgements which have to be made when, for example, data such as buckling coefficients are being evaluated for incorporation in a data base. Data which have been deposited in such a base require to have specified alongside them a statistical quality, a tolerance on accuracy or a limitation as to the circumstances under which they may be applied. Some of these limitations may be very straightforward. In the case of the example we have been considering above it is likely that a limitation to the elastic range of the material will need to be placarded; this would be a consequence of theoretical considerations. In the same example it would also be necessary to specify the range of types and configurations of edge support within which the data are likely to be reliable. This would quite often be a consequence of more subjective judgements based upon expert practical knowledge and experience.

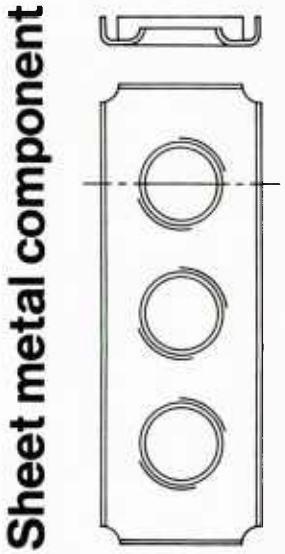
Here too is an important consequence of the means by which such data are stored and communicated to the design engineer. Data banks in printed form, such as manuals and data sheets, offer considerable opportunity to elucidate the quality of data and to specify that quality in numerical terms and also in more subjective terms. The electronic forms of these data are currently limited to the representation of numerically specified limitations which are not affected by subjective influence. We

are only just beginning to get to grips with the possibilities of "knowledge based systems", or "expert systems", and the opportunities they may offer to incorporate into our data bases information which will reflect the subjective influences which may have been applied during their evaluation.

To illustrate that we are here dealing with questions of considerable importance to the health of our industries, and to our ability to produce cost effective products for our consumption and defence, we now turn to some practical examples.

COST AND OTHER CONSEQUENCES OF INADEQUATELY REFINED DATA

The industries of the free world are increasingly competent and surprise most of us with their ability to provide increasingly elaborate goods and weapons to fill the ever more sophisticated desires and needs of the society they serve. They do well enough, one might think, with the data resources already to hand even though the quality of those data may not be as high as many of us have reason to believe it ought to be. One of the reasons is that there are substantial benefits to be gained by increasing the quality of the data which are available and making their storage and retrieval ever more convenient and reliable. One of these benefits is in cost saving and I will illustrate this with an example I have used many times before but which makes the point most effectively. It also illustrates the variability of the quality of data in use even in our largest and best informed companies.



Estimate by:	Buckling strength in shear:
Company A	0·79 X
Company B	1·07 X
Company C	1·63 X
Company D	0·94 X
Company E	0·72 X
Solid, simply supported plate theory	X

FIGURE 2 VARIABILITY OF DESIGN DATA IN USE

Some years ago I undertook an exercise to compare the data in use by 5 different, quite large, companies relating to just one common design problem. It is illustrated in Fig. 2. The data actually being used by the 5 companies for estimating the buckling strength in shear of sheet metal components containing flanged lightening holes were collected. They were then applied in turn to a typical design of specified dimensions and material. The strength estimated by the 5 sets of data varied widely; there was a 2 to 1 ratio between the greatest and least strength predicted. Yet, as far as was known, no premature failures had been recorded for any actual design based on any of these data; almost certainly some of the design teams using these data were overdesigning. The type of component involved is very common in lightweight structures, including those of transport aeroplanes. To illustrate some of the cost consequences of this variation an estimate was made of the differences, due to weight, in the revenue earning potential of two hypothetical medium-sized transport designs which were identical except that one would utilise the heaviest and the other the lightest components designed in accordance with the extremes of the data in use. For a fleet of 10 aircraft, the difference in revenue earning capability was found to be close to $100,000 per year in present day values.

Such disparities in the quality of data used by different organisations are rarely revealed - at least publicly. There are circumstances which cause such disparities to come to light at least to a limited audience. Companies in the western world have, for sound economic or political reasons, found themselves in collaborative design and

production arrangements particularly in the military field. These arrangements are never easy to manage and they are not helped by finding that the basic design data in use by one partner differ significantly from those in use by others. Much time and expense, let alone goodwill, may have to be sacrificed before an agreed common data set can be established. Disparate data resources can also have important commercial consequences. Large companies seeking tenders from sub-contractors may be led astray, or at least be put to much trouble, if the sub-contractors concerned are basing their offers on entirely different basic data from those which the principal assumes as valid; such a circumstance may not even come to light until the final commissioning of a piece of plant or machinery.

Data of inadequate quality may often mask an opportunity to employ less expensive production methods. For example, data relating to fatigue strength, surface finish and geometric stress concentration can be used to make decisions on quite minor geometrical changes in machine components which will permit turned finishes to be employed rather than more expensive ground finishes yet achieve a comparable fatigue life. Decisions such as these, however, can only be valid if the data used have been properly evaluated; small inaccuracies and misunderstanding of limitations could lead to erroneous conclusions.

In many engineering problems there is a domino effect which can multiply the potential of a component to cause catastrophe and loss out of all proportion to its superficial importance. Establishing uniformity of design quality and therefore of the quality of data used cannot be confined just to the overall design. When trouble occurs it is frequently initiated at a local 'hot spot' which is too small to be detected by many of the analytical techniques. Similarly a train of events leading to catastrophic failure may be initiated by a mechanical component which normally plays but a minor role. Large capital equipment such as aircraft and chemical plant are particularly vulnerable in this regard. Failures are not always catastrophic; more often they simply eat away at operating costs and company profits. For example, a case is known where a single-stream chemical plant suffered from a number of stoppages due to repeated failures in a shaft which was part of a pump in the auxiliary equipment. The loss of production due to shutdown while repairs were effected was valued at $240,000 per day. These losses continued from time to time and appear to have been accepted almost as inevitable until it was identified that a stress-concentration effect had been wrongly evaluated. Better quality data were brought to bear in a redesign and no further losses were experienced.

Of course, the cost consequences of such failures can be astronomical and there are other consequences which the professional engineer cannot afford to ignore if he is to retain respect and influence in society. Ernest Breton (7) references an attorney's estimate that 25% of the lawsuits brought against new products allege engineering negligence. He also recalls the testimony before the President's Commission investigating the Three Mile Island nuclear accident. Among other problems it was alleged that the facility was plagued by malfunctioning valves, and unreliable instruments. The engineer is particularly vulnerable, compared with other professions, in that the overall results of his work can generally be assessed by objective means and by repeatable experiment. In general terms attorneys, and society at large, may acknowledge that perfection is never possible in the practical world. But this is of little comfort to the engineer called upon to justify the many imperfections which invariably come to light during the investigation of a failure even though those imperfections have not contributed to the failure and, in some cases, may even have been guarded against in a fail-safe system.

The only protection the engineer can obtain is that which he provides for himself. This was, perhaps, more easily done when every stage of the design was under his direct surveillance than it is today when so much may be hidden from him in a computerised system. The increasing use of these systems places an added responsibility on engineering management to be fully aware of the principles on which their programs are based and, particularly, of the quality of the data which they draw upon. But it does go further than this; when data are embedded in a computerised system they need to be of more certain quality than would be the case if they were more readily available for routine inspection. A given data value in computer memory may be accessed for use in a number of different applications. In some applications an error of substantial magnitude may have little effect but be quite serious in others. The only course open is to ensure that data of a quality appropriate to the most sensitive case conceivable are stored. The possibility open to the human operator to seek a high quality data value only when its quality is particularly critical may no longer be available.

REASONS FOR VARIATIONS BETWEEN DATA SETS

When a data base is being constructed we start with a vast collection of raw material. This raw material is frequently in the form of reports originating from basic research and from applied research which has been undertaken during the development of products. It may also originate from tests of the actual performance of finished products. These records are generally numerous and their content will include commentaries on how the data were derived. If they are experimentally derived the report will, in addition to providing numerical results, contain more or less extensive commentaries on such things as the design of the experiment, apparatus used, the test methods employed, the readings taken, corrections made to these readings together with ancillary readings of such things as ambient temperatures and pressures to support corrections and much else.

If the data are derived by theoretical prediction methods the position is similar except that in this case assumptions will be stated, idealisations will be described in order to make the problem amenable to mathematical modelling and so forth. The resource of raw material from which a data bank may be constructed contains much else beyond the raw numerical values.

At first sight collections of raw material like those described above may seem to consist of two parts, data of interest in an engineering application and 'noise'. Indeed, it often seems to the engineer that the research worker, experimental or theoretical, is more concerned with scientific methods or mathematical principles than with the practical value of the results they provide. But any temptation simply to transfer data into a data bank and ignore the other information provided has to be resisted. The fact is that there may be little consistency in the information provided in support of experimental results between one researcher and another. This is particularly common in relation to corrections made, the detailed circumstances of the test and auxiliary information recorded. For example, though there has since been much improvement, in the early days of fatigue testing it was not unknown for such 'details' as whether a fluctuating loading had been superimposed on a steady loading to escape mention or for the frequency of application of the fluctuating loading to be unrecorded. The first of these omissions invariably is serious and the omission of frequency may or may not be serious depending upon the material of the test specimen. This suggests other features leading to variations in raw data sets. As time progresses experimental techniques develop and new parameters are identified which can affect the results obtained.

Particularly in relation to data which have been generated from research undertaken to support product development there is much scope for 'error' or, to put it at its best, inconsistency. The researcher may himself be unaware of these inconsistencies. They may never come to light unless there are collected together several raw data sets on nominally the same phenomenon which exhibit a cloud of points rather than the neat relationship which might have been expected or which a mathematical model of the phenomenon has been found to predict. This may lead the experimental man to conclude that the theory is wrong or the theoretician to conclude that the experiment is wrong. More often, I suspect, neither is wrong. It is more likely a question of inconsistencies in the assumptions made, the idealisations employed and the corrections which have been applied.

Within this vast untidy resource is the raw material from which we construct data banks; the raw data resource does not of itself constitute a data bank upon which valid engineering decisions can be based. In its raw form it may present a picture only of conflict and confusion. Part of the methodology in resolving the situation consists first in recognising that this is the case and to assist our appreciation we may summarise those characteristics of the resource which require our special attention. Raw data are:-

- Often vast in number
- Widely scattered
- Lacking consistency, both in terms of quality and back-up information
- Prone to error and bias
- Variable in respect of time

In respect of the last characteristic two points should be made at this stage. First of all I do not subscribe to the view that information of greater than a certain age necessarily has little value. Engineering problems have a habit of returning in response to economic needs or even the whim of fashion. For example, for economic and other reasons propeller driven aircraft have started to come back into vogue in several missions. We found, in my own organisation, that before we could proceed to evaluate propeller data appropriate to modern configurations we had to start with data of some considerable antiquity and to make reference to experts some of whom were near retirement. It was a foundation which could not be ignored though it was difficult to establish.

The second point which has been made relates to the observation that experimental and theoretical techniques develop with time. This means that it is essential that data banks be subject to continuing review and updating as necessary. In this regard, I have for many years advised engineers to apply the following simple test from time to time. Speculate that a particular numerical data value being used may have drifted by 1% for each year since the time it was last evaluated or since the time the company data manual was last updated. If the effect of such a drift on the predicted performance of the product they are designing is of significance then it is advisable to have the data value reconfirmed. This is a simple rule-of-thumb but it is recommended by several different experiences of the extent to which even the most carefully constructed data sets may be overtaken by better knowledge.

MANAGEMENT OF EVALUATION AND VALIDATION - PERSONNEL

Up to this point we have examined, amongst other things, the types of data with which we are concerned, the circumstances under which they are applied in industry, the circumstances under which raw data are generated, some of the qualities of the raw data resource and those needed in its refinement. This excursion has been necessary because, before discussing how a task may be accomplished, we have needed to observe who will benefit from our efforts, what are their expectations and who else is involved. As with most management problems we find that we are eventually dealing with people and our first purpose is to identify who they are. In addition we need to keep in mind the cost of what we are doing and whether those costs will be supported by those who will benefit from the data bases we construct.

We immediately face a dilemma. If the data base we construct is made of sufficient scope to provide data in response to any demand, even within a fairly limited discipline, it is unlikely to be economically viable. The price paid for the benefit of those data which are requested will also have to support the costs of producing and refining data which are never requested. At the other extreme if we limit the base only to those data within a field which we can identify as being in most common use we may sterotype the designer's work and inhibit progress. My own organisation has had to steer a very careful course between these extremes and has from time to time been open to criticism on both counts.

In the absence of a sufficiently generous benefactor, I have found but one course open to provide a workable compromise. One group of people who must be involved in the system is a representative group of eventual users and another is a representative group from academia. The user has to be involved in order to specify the sets of data, in a given field, in which he is currently interested or for which he can perceive a need. These will be heavily biased towards data relating to his current problems. To balance this, the academic is introduced since he is able to flavour the scope addressed with a reasoned basis for including data relating to developments which are likely to have attractive futures. He is also very often able to warn against a search for knowledge which does not exist.

Having decided the personnel needed to be involved at the stage of setting the scope of the data base, and the review of that scope at appropriate intervals, we now turn to the processes of evaluation and validation proper. In relation particularly to phenomenological data we have noted, in a previous section, the importance of the circumstances of the application of data in regard to such things as the accuracy required. Here again, representative users of the data have an important role to play. We have already noted the possible dependence of data upon subjective judgements which can only be made by those having appropriate skill and expertise. The extent to which the generators of data would themselves admit to the presence of subjective influences upon their work, even within the terms of this discussion, may not be very great. They may, however, be more ready to evaluate the effect of these influences on others! But in any case only the generators of raw data are in any real position to draw attention to those features which may not be accounted for by well recorded, repeatable means.

Accordingly, the personnel involved must include representative groups, albeit quite small, of industrial users, academics, and data generators. It should be noted, of course, that individuals from any of these three backgrounds may in fact span more than one of them. Many academics are themselves data generators for example.

A group constituted as I have so far described it needs further elaboration. Data evaluation is a time consuming operation. Practising engineers, academics and raw data generators have neither the time nor often the inclination to be involved in what they might see as the drudgery of data collection, evaluation and refinement which follows the stage when a well specified need for data has been drawn up. Beyond this, however, the environment in which these people work does not fit them ideally for this. Reference 8 examines this environment and changes taking place in more detail.

In the part of the task where raw data are collected, analysed and distilled down to the best set for application to a specified need, we require special staff of flexible outlook. They need a good academic background in the discipline to which the data they are handling relate and preferably have a short working experience in either research or design. Special training beyond that obtained in their academic careers must be given for several purposes. First the techniques of managing the various interests which will be at play during the data refinement process must be imparted; secondly the procedures of data refinement itself must be learned.

Having identified the need for a particular data set, the raw data appropriate to this set have to be identified and collected. Libraries, bibliographic retrieval services and all other means available to collect as much of the published raw data as possible are used by the specially trained staff. But beyond this, on most topics, there is a body of raw data which is <u>not</u> published; it is not uncommon for this to be as much as 50% of the total in existence. It can be released from the files of companies and individuals but only by sensitive management to obtain their co-operation and interest.

Once the numerical data have been extracted they will usually be found to present a conflicting and confusing picture. Digging into the circumstances under which experimental data were derived, making a full technical assessment involving available theoretical treatments, or involving the development of new mathematical models, are just some of the time consuming and expensive processes which have to be undertaken if the preliminary work of data refinement is to be sound. It isn't just a matter of putting a 'best fit' curve through a cloud of data points. Even when the raw data appear not to be in conflict the simplest processes of combining the data of several sources can lead to surprising results.

The staffing and processes of data refinement need to be undertaken in an environment which is, as far as can be assured in today's world, truly impartial. This does not exclude, in my view, the environment of a commercial company, or of a research organisation or of academia if one can be assured that there is no possibility of a vested interest in a particular data set being present. Clearly it is most easy to satisfy that condition in a body which is not itself involved in research or raw data generation.

Throughout the initial process of drawing together a preliminary refined data set, the staff person will have access to the industrial users, academics and data generators in the group which he or she is serving. Many of the techniques which other papers in this series describe in some detail will be applied. Here we are concerned more with data the evaluation and validation of which are influenced, at least in part, by subjective influences and most particularly by the skill and experience of the group we have identified in the previous sections.

We may now summarise the set-up we have created. The first objective we have so far identified is to define the scope of the data base we are to create within reasonable economic parameters. To achieve this we require the involvement of an expert group including representative users of the data and of generators of the raw data which are to be evaluated in the process of building up the data base. We also have seen the advantage of including academics. For the purpose of identifying, collecting and correlating the raw data we have identified the need for specially trained staff. In the modern application of this system, of course, their work of identifying and collecting raw data will be assisted by on-line bibliographic retrieval and the employment of information specialists. The special staff will mostly be engaged in the correlation of data and its evaluation by well-established objective principles and will draw upon the knowledge available to them amongst the individuals in the expert group. Using this mechanism we can arrive, fairly surely, at what we will call a first-draft representation of a set of data in answer to a perceived need. How these personnel are to be organised to complete the evaluation of the data set, and validate its use in practice, we shall next consider.

MANAGEMENT OF EVALUATION AND VALIDATION - ORGANISATION AND METHODOLOGY

Once the process of data refinement, against some well specified need, is under way and in the hands of trained staff it is necessary to make arrangements for the validation of the refined data packages which are being produced. Having noted some of the pitfalls in data correlation we may well wonder if the data produced by even the most wary, qualified and independent person can be trusted. We must bear in mind the potentially high cost of failure in many of today's larger engineering enterprises, the increasing social pressures in connection with the avoidance of environmental and other accidents not to mention the increasing extent to which manufacturers and designers labour under the spectre of liability suits. Contemplation of this question very early in the life of my own organisation led to the conclusion that data distilled from the mass of sources must be monitored as they are produced and not issued for use until they have been validated.

In 1977, when Dr Frank Press appeared before the Senate Committee on Commerce, Science and Transportation prior to his induction as Director of the Office of Science and Technology his reaction was sought to the suggestion "that there was no such thing as objective technical advice". In his answer he rightly observed that the best an individual such as himself could do would be to present the known biases along with his advice. That probably is the best that an honest individual can do - what I would suggest, and what my experience bears out, is that the power and reliability of individual expert judgement can be magnified many times when applied as part of a carefully managed consensus seeking group. The application of such a group is what has to take place in the process of validation.

"Consensus seeking group" will immediately equate, in most minds, with the term "committee". This is in some ways unfortunate and, I must admit, such groups employed in my own organisation are referred to as committees or steering groups. Committees are used throughout the world for many purposes some of which only have a veneer of seeking true consensus of opinion. They often, for example, consist of variously qualified individuals, each representing an interest and determining issues by some system of majority voting. But such is not a true consensus group and the inherent potential unreliability of such groups was exposed in some detail long ago by Charles Dodgson (Lewis Carroll) (9). In our present case, it would not suffice to have issues on the validity of our data sets qualified by some majority voting procedure.

In the absence of a staff member of the group, who has a quite different part to play from that of the individual expert, we would find the exposure of technical issues and their rigorous examination inhibited by personal relationships. For example, where one expert member of a technical group is asked, (or more usually is persuaded to volunteer), to undertake a technical task and report back to a committee of his peers, psychological and social forces can operate so as to inhibit the subsequent rigorous discussion and judgement of that task. On the one hand the other members of the committee may restrain their comments lest the resolution of an unsatisfactory task should be that it is transferred to them! On the other hand, more rigorous comment may imply serious criticism of the voluntary expert's professional status and integrity. In my experience of attempts of this sort the latter is the greater danger and a debate ensues in which the chosen expert's subsequent defence of his professional reputation detracts from the purposeful achievement of consensus on a technical issue. The most effective procedure is to have, as part of the structure, one member who is appointed to serve the group's technical requirements, who is sufficiently well qualified to appreciate the issues but who does not associate his or her professional status with an eminent position in the field with which the group is concerned. This, of course, is the specially trained staff member whom we have previously included for quite different reasons.

So the basic organisation in the management structure comprises a group, which for convenience I shall continue to call a 'committee' bearing in mind the observations made on this term in the foregoing paragraphs. With this committee there is closely associated a staff engineer. We have already described the duties of this staff engineer up to a point of preparing a 'first draft' data set. By what methods are evaluation and validation completed by the committee organism?

Of the several tasks to which the group has to address itself, being assured on such as the following questions are among the most important:

Has all reasonable care been taken to find relevant data; have all sources known to the wider group of experts been tapped?

Are the correlation processes used sound; have any philosophical or mathematical traps been fallen into?

Are the necessary limitations on ranges of applicability of the data and other cautions specified?

Is the presentation which has been used clear and convenient remembering that application of data is made by engineers not, in general, by scientists?

Each expert has to be prepared, individually, to support the contention that the refined data package represents the best data on the topic in question, at the present time and within the limitations specified.

A large proportion of the issues which arise in such a committee can be resolved by objective methods. These methods can be specified by the expert component of the committee and the staff member discharged to carry them out as part of the preparation of a 'second draft' data set. Some issues defeat strictly objective examination. Judgements upon the validity of data, and the specification of the limitations within which they may be applied, can often rest upon skill, experience and factors which are more subjective. Examples of such issues include the extent to which parameters, not recorded for a set of test results, may have affected the data. Another may be the extent to which an idealisation in a set of experiments or in a mathematical model may be acceptable as representative of the practical application. Another issue may be whether there are empirical corrections which can be devised; these may be based perhaps on little more than general experience of the effects of parameters which are not explicitly defined but which may reasonably be assumed. Yet another may be the extent to which an approximation representing a given regime of behaviour may safely, within a specified range of application, be extrapolated into an adjoining regime. A dramatic example of this last type is commonly employed and accepted in structural design. Most analysis is still conducted on the basis that metallic structures behave as though their components were elastic up to failure. Though it is known that this is not the case experience dictates that it is an acceptable basis for design; paradoxically it is the very fact that these materials are not elastic up to failure which permits yielding at stress 'hot spots' with subsequent favourable redistribution of load elsewhere! The expert also knows, in this case, that in particular applications the user of the data will be constrained to apply factors of safety which take accont of these effects - but great care and judgement are nonetheless required.

On matters of judgement on phenomenological data, I believe the constitution and organisation of a management system along the lines described is essential. Its central feature is the achievement of consensus of opinion on the judgements which are made in relation to all matters of substance. In turn, the judgement of what is 'a matter of substance' may itself be largely subjective.

A SIMPLE MODEL OF THE PROCESS OF CONSENSUS

I mentioned the amplification in power of the individual when applied as part of a

consensus seeking group. By consensus I mean what the lexicographers mean by consensus; unanimous and not majority agreement. Obtaining consensus is often a long and expensive business - but well worth it as I hope now to demonstrate with the aid of a simple model.

Let it be supposed that a series of technical issues can be framed as questions to which we happen to know, by some test or other, that there is a RIGHT answer or a WRONG answer and none other. If you were to pose these questions to one expert he might take the RIGHT decision on, say, 80% of occasions being WRONG on the other occasions because of erroneous thinking, lack of experience or bias. Quite separately we now take another expert, equally well qualified and prone to error and bias. So he too would select the RIGHT decision on 80% of the questions posed and the WRONG one on 20% of them. On average over a large number of such tests we would find that for every 100 questions posed:-

The two experts gave the same RIGHT answer (consensus) 64 times
The two experts gave the same WRONG answer (consensus) 4 times
One chose RIGHT and the other WRONG (no consensus) 32 times

If we accept only those answers on which there was consensus, we obtain 64 RIGHT answers for every 4 WRONG answers or 16 to 1.

Thus a consensus group of two experts increases the reliability of the decision taken by an individual from 80% (or 4 RIGHT answers for every WRONG answer) to about 94% (or 16 RIGHT for every 1 WRONG).

What happens if we extend this model to a larger group? With 6 equally 'reliable' individuals in a consensus situation the reliability increases from the individuals score of 4 RIGHT to over 4,000 RIGHT for every one WRONG.

There is, of course, a price to be paid. It will have been noticed that of the 100 issues addressed to the group of 2, there were 32 occasions on which there was not consensus. In practice what has to happen is that the issues are rephrased again and again, by the staff serving the consensus seeking group, until there are sufficient features on which unambiguous agreement can be obtained. The simple model reveals that with a 6 member consensus group one would have to phrase nearly 400 questions to obtain 100 on which consensus would occur. This would take 4 times as long as posing 100 questions, which is the number necessary to get 100 'consensus' answers from an individual. Is this expansion in the time taken worth it? I think the answer has to be in the affirmative. The group of 6 was RIGHT 4,000 times for each WRONG whereas the individual was RIGHT only 4 times for each time WRONG. Taking 4 times as long to achieve a reliability increased by a factor of 1,000 must be worth it in the circumstances of modern technological endeavour. The time taken to reach consensus is increased for tasks in which the reliability level of individual participants is lower than it is for other tasks. Thus, agreement on data sets towards the boundaries of an expert group's experience takes much longer in the consensus process than do more familiar sets. This is quite apart from whether the processes of finding and correlating data prior to the validation stage has been more time consuming in such cases.

Another consequence of this "simplified theory of consensus", is that the reliability of the validation decisions taken by such a group is disproportionately weakened, as one might expect, by the inclusion in the group of individuals whose own reliability is low due to inexperience, extreme bias or whatever. What one might not as quickly recognise, though it is overwhelmingly the general experience of anyone who has sat on any sort of committee, is that the inclusion of but one such biased member can extend the time taken to reach decisions to an alarming degree!

Of course, this model is a considerable simplification of the circumstances obtaining in an actual group of experts constituted in a committee situation. Amongst other things it ignores the influence which one member may have on the judgement of another. Here again is a pointer to the type of expert whom it is ideal to have involved for they should not too easily be influenced by the judgements of others but use their own experience and skill to evaluate the judgements of their peers. The model has also been based on the case where there is a RIGHT answer, a WRONG answer and none other. In reality whereas there may be one RIGHT answer there may exist a universe of WRONG answers. In that case the risk of consensus on a unique WRONG answer is much lower and the overall reliability of the process is still further enhanced.

Despite its limitations, however, the simple model is remarkably revealing of features which I have had considerable opportunity to observe, in committee situations, over many years. We have examined, with the model, the question of whether the consensus process is efficient in reliability terms. What of the costs which are involved in the evaluation and validation of phenomenological data?

COMMENTS ON COSTS AND VALUES

Fig. 3 presents a histogram of the relative frequency of data evaluation tasks, as a function of their relative costs, covering a decade of activity within my own

organisation. It is worth observing that the mean cost is at about 3.8 on the scale of Fig. 3. To a large extent cost and time are interchangeable. The mean time taken to reach a 'first draft' quality by the staff man working alone would be at about 1 on the same scale. The mean elapsed time to completion of 3.8 times this value gives some indication of the very real input which the expert committees induce. It would not be correct to assume that the staff man would regard the quality of the data at the 'first draft' stage as satisfactory. Nonetheless, if he were working alone it is unlikely that he would consider his work but some quarter completed by the time he reached that stage. It is beyond our present scope to apply the model of the consensus process to these results though we may observe that the general features are consistent. The long tail of the distribution shown in Fig. 3 is generally influenced by the time taken to reach consensus on tasks which are novel or towards the boundaries of an expert group's experience.

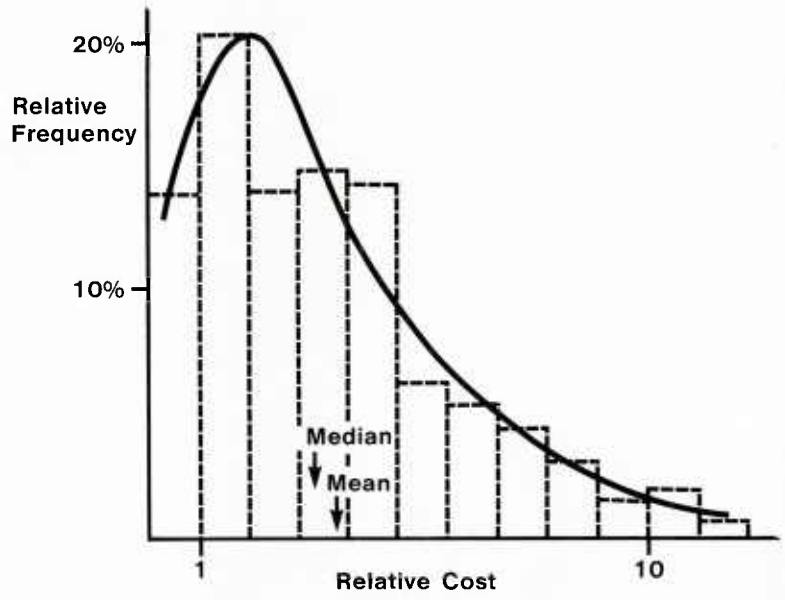


FIGURE 3. FREQUENCY OF OCCURRENCE OF DATA EVALUATION TASKS OF DIFFERENT COMPLEXITIES

The histogram of Fig. 3 is skew. Data evaluation and validation is not a production process which it is easy to schedule. Experience of the time taken in preparing one data set, of the type to which our present discussion relates, rarely reads over into other data sets. As with research, it is difficult to forecast the time which will be taken to reach a point where no-one has ever been before and where even the direction in which to go may not be particularly clear at the outset! The long tail of this distribution is understandably daunting and frustrating to those of us responsible for planning, budgeting and marketing any numerical and factual data bank activity.

Whilst it may be possible, as in my own organisation, to collect historical cost data and even to project costs over a reasonable agglomeration of data preparation tasks, it is almost impossible to relate these to the value which the completed data set will have in application. In an earlier section of the paper it was possible to put monetary values on the costs of inadequate data which led to a failure or demonstrable inefficiency of some kind. Evaluated data, hopefully, do not give rise to failures and direct value benefits cannot be demonstrated in as convincing a way.

In general terms, data users will agree the broad features illustrated in Fig. 4 which suggests economic considerations which should not be overlooked.

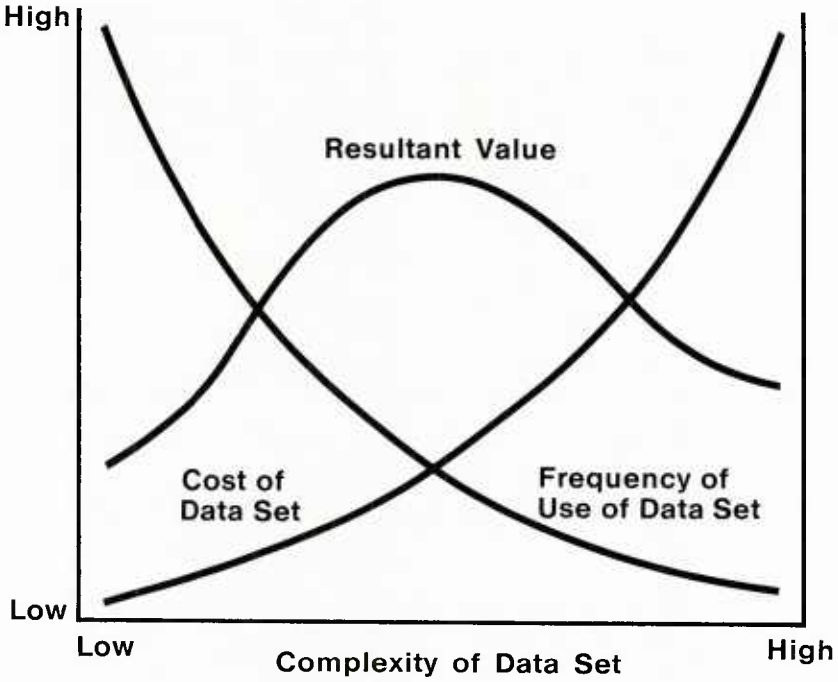


FIGURE 4. COST AND FREQUENCY OF USE OF DATA SETS

The two basic curves of "Cost" and "Frequency of Use" are compatible with the features of the histogram of Fig. 3 if one assumes that the tasks most frequently undertaken in my organisation are those relating to data most frequently used in engineering practice of the disciplines to which they relate. This is a valid assumption.

Fig. 3 shows that data sets of lowest relative cost are those most frequently applied in engineering practice. The most costly tasks generally correspond to data of higher complexity, that is, those which are of more specialist interest and therefore are least frequently applied, in relative terms. If the price which a user pays is equitably based on cost of production then the resultant value perceived in a data set is some function of the product of cost and frequency of use. The highest 'value' is therefore placed, in the general terms we are using, on data sets of some intermediate degree of complexity. This has several implications for anyone embarking on the construction of a data base whether it is undertaken for commercial reasons or whether the data base has to demonstrate its relevance to national need and justify the expenditure of public money.

Fig. 4 provides some suggestion of the trends which might be expected in the value placed on refined data with the growing expectation of data banks being accessed more frequently by the Computer Aided Engineering (CAE) phase of Computer Aided Design (CAD). Further comments on the economics of data bases are to be found in Ref. 11.

THE FUTURE

Numerical data banks are being developed increasingly for the benefit of CAE. The computer offers the opportunity to try more variants quickly at the design stage and, in the terms of Fig. 4, the frequency of use curve will move bodily upwards. This moves the optimum on the resultant value curve to the right suggesting that greater value will be perceived in data sets of higher relative complexity. This assumes a static cost curve. It may be that, because of features noted in previous sections of this paper, a generally higher degree of refinement and reliability will be demanded in all data sets which are embedded in computerised systems; I would like to think so. In this case the cost curve moves bodily upwards also and the position of the optimum in value terms is restored to the left. What is more important is that, in both cases, the resultant values are increased at all degrees of data set complexity.

The new generations of computer systems offer the ability to apply so-called "artificial intelligence" in expert systems. These systems enable the manipulation and automated application of factual information much of which is not expressible in numerical terms. It is particularly in relation to factual information which is not expressible solely in numerical terms that the philosphies and methodology described in this lecture will have most particular relevance. Whereas much numerical data can, by its nature, be evaluated by objective means, non-numerical factual information is more likely to be susceptible to less tractable skill and experience judgements and considerable impetus for the development of methodologies to deal with these can be foreseen. It will certainly be needed for I would not represent those discussed in

this lecture to be other than a starting point even though they have developed from origins going back at least 40 years. Speaking at a conference in London a few months ago, Max Bramer, an authority on expert systems, is reported as claiming (10) that 'experts are always wrong' and 'that it follows that all expert systems are wrong too'. Experts certainly make mistakes but, in relation to the type of application and the types of expertise which have been described in this paper, I hope it is clear that the subjective influences which they can bring to bear have a value which is irreplaceable and, through careful management, can be dramatically enhanced.

REFERENCES

1. LOCATI, L., Scientific Foundation for Development of Modern Industry, 24th Meeting of AGARD Structures and Materials Panel, Turin, 1967.

2. BARRETT, ANTHONY J., Basic Principles to be Observed in Preparing Evaluated Data for Industry, 2nd CODATA Conference, St Andrews, Scotland, 1970.

3. COX, LESLIE H., The Design of Structures of Least Weight, Pergamon Press, Oxford, 1965.

4. ESDU, Drag Rise Mach Number of Aerofoils Having a Specified Form of Upper-Surface Pressure Distribution: Charts and Comments on Design. Engineering Sciences Transonic Data Memorandum 71019, ESDU, London, 1971.

5. MIL-HDBK-5C, Military Standardization Handbook, Metallic Materials and Elements for Aerospace Vehicle Structures, Department of Defense, Washington D.C., September 1976 and later issues.

6. ESDU, Metallic Materials Data Handbook, AvP 932, ESDU, London, 1981 and later issues.

7. BRETON, ERNEST J., Why Engineers Don't Use Databases, ASIS Bulletin, August 1981, John Wiley & Sons.

8. BARRETT, ANTHONY J., Refining Data Resources to Assist Technology Transfer, Technology Transfer in Industrialised Countries, Sherman Gee (Ed.), Sijthoff & Noordhoff, Netherlands, 1979.

9. DODGSON, CHARLES L., Pamphlets on Committees. Reproduced in Theory of Committees and Elections by Duncan Black, Cambridge University Press, 1968.

10. BRAMER, MAX, 'Expert Systems Always Wrong', Report by Philip Hunter, Computer Weekly, 25th November 1982.

11. WESTBROOK, J.H., and RUMBLE, J.R., (Eds), Computerised Materials Data Systems, Steering Committee of the Computerised Materials Data Workshop, Fairfield Glade, Tennessee, 1982.

The Evaluation/Validation Process

Data from Disciplines Resting on Good Theoretical Foundations

Dr. Malcolm W. Chase
The Dow Chemical Company
Thermal Group, 1707 Building
Midland, Michigan 48640 USA

SUMMARY

The concurrent use of experimental and theoretical techniques can be used to solve problems efficiently. To this end, however, the techniques must be used with a well-defined plan in mind. A literature survey is first required to reveal available information which relates to the problem. The meshing of this information and any other soon-to-be-released data is an important second step. The experiments and/or theoretical calculations must then be coordinated to reduce time and expense and to provide maximum data at minimum expense. From our experience in developing the JANAF Thermochemical Tables, examples will be given where calculations have been of sufficient accuracy to reduce the need for experimentation as in the thermodynamic properties of some chemical species. Comparisons and trends in data are also valuable to extend or replace data. Four thermodynamic studies will be used to illustrate the value of theoretical efforts.

INTRODUCTION

As in my previous talk, this presentation will discuss the experiences of a laboratory in its never-ending search for data and its use thereof. As mentioned earlier, in order to perform our activities efficiently, our experimental and theoretical efforts are designed to build on existing information. In addition, to be efficient in terms of time and money, an evaluation between experimental and theoretical efforts is made as to which approach or combination thereof is satisfactory for the current problem. The theoretical effort oftentimes is treated with disrespect; the feeling by some is that the only reliable result is an experimental result. This attitude may be more prevalent in the industrial environment than in the private/university laboratories. In some industrial applications, however, government regulations (at least in the United States) often preclude a reliable thorough theoretical effort.

This presentation will discuss the two major contributions of theoretical endeavors in scientific projects. First, theoretical approaches are used to provide data; data which is presumably not currently available from experimental measurements. This approach is often treated by skepticism. Second, theoretical approaches are used to analyze the data and relate it to the final results of some specified project. This second approach also provides data (perhaps your final result) from disciplines resting on good theoretical foundations. To illustrate the usefulness of such efforts, a modification of a presentation by Joseph R. Downey, Jr., on the role of the data design laboratory will form the larger part of this discussion.

The data design laboratory occupies a unique position in the chemical and allied industries. This laboratory shall be defined for our purposes as one which supplies data that forms the basis for process design calculations. In the chemical industry these process design calculations are normally made by process engineers using physical property data, miniplant/pilot plant experience, and a variety of calculational techniques including sophisticated process simulation programs such as ASPEN PLUS†. The results of these calculations can be no more accurate than the physical property/phase equilibria data that was their basis, no matter how sophisticated the simulation procedure may become. The successful (safe, efficient and reliable) plant design requires an accurate physical property data base coupled with pilot plant experience and reliable process simulation calculations. The trend in the industry in recent years has been to reduce or eliminate the costly pilot plant and/or miniplant stages whenever practical in order to have production plants on stream faster and at lower overall cost. This places an even greater burden on the reliability of the physical property data and process simulation calculations. This presentation shall focus on the role of the data design laboratoy in providing reliable data for these purposes. The importance of accurate thermophysical property data for a number of industrial applications has been reviewed by Sengers and Klein (1).

Some of the types of data frequently required for process design purposes are listed in Table 1. This is not an all-inclusive list but is intended to represent the most commonly required types of data. The data design laboratory ideally should be equipped to supply these types of data under all required process conditions. In practice, experimental attainment of this goal is impractical because of the very diverse conditions of temperature and pressure required for some specialized processes. The approach generally adopted by most companies is to set up one or more laboratories equipped to supply the types of data listed in Table 1 over conditions of temperature and pressure commonly encountered in that company's processes. These conditions may

vary significantly from company to company due to the nature of the products produced. For example, a company specializing in cryogenic materials may have rather different process data requirements than a company which never deals with cryogenic materials.

In order to achieve its maximum effectiveness the data design laboratory should be far more than simply a laboratory to perform experimental measurements. Instead it should encompass all of the following capabilities:

1. literature awareness
2. calculational and estimational methods
3. experimental methods
4. critical evaluation

When configured in this manner, the data design laboratory can work with the process engineers to determine (a) what data is already available in the literature that bears on the problem at hand, (b) what additional data can be estimated or calculated, (c) whether any or all of the above data is of sufficient reliability for the current use, and (d) what additional experimental or theoretical data is needed and how can it be obtained. Each of these capabilities will be discussed separately in the following sections but the real value of the data design laboratory is the interplay of these activities to solve the problem at hand. The combination of these capabilities in one laboratory results in a synergism which would not be present if they were scattered into several laboratories. For any given problem, one of these capabilities may prove to be more important than the others but, in general, this will not be obvious until the problem is treated in detail.

Also of importance is the impact of the data design laboratory on problems other than those relating to process engineering data needs. When configured in the manner described above with emphasis on the areas of thermodynamics/physical properties/phase equilibria of traditional interest for process engineering data needs, such a laboratory is also ideally suited to a number of other data needs within the company. Most notable among these are the reactive chemicals area and the estimation or measurement of properties required for product specifications or for a variety of Research and Development or Technical Service and Development applications.

LITERATURE AWARENESS

This area includes a knowledge of typical sources of data or bibliographies leading to data as well as a general feel for new or improved theoretical or experimental techniques. One of the most important aspects is awareness of information sources. These sources are the key to the work which has already been done. The primary scientific literature is the main data source and must be searched thoroughly and completely for each specific piece of data of interest. This search can be quite lengthy and costly if not done by a skilled searcher. Typically this takes less time and less money than the actual new measurement or calculation. In some cases the literature search may consume more time and effort than new measurements would have taken but this is unusual and typically occurs only if the focus of the search was very narrow; e.g. specific gravity of a single material at 25°C. Broader literature searches on multiple properties of a given material or one property for a variety of materials are very rewarding. The results of such searches form a data base which is valuable input to the process of deciding whether to estimate or measure missing or suspect data.

Searching of the scientific literature is appreciably eased by a number of secondary sources devoted to specialized areas. The disadvantages of these secondary sources are that it may be difficult to determine how thoroughly they cover the field of interest, and they will normally lag or be out of date with respect to the primary literature. Nevertheless, they serve a very useful purpose and the thermophysical property area is blessed with a number of them which greatly simplifies searching of the literature. These secondary sources range from bibliographies, to tabulations of all experimental data, to tabulated data with some correlation to commonly-used equations, to tabulations of selected values, to critically evaluated compilations. Each type has its own advantages and disadvantages which shall not be enumerated here. What is highly important is that the researcher understand which type of source he is dealing with and not mistake one for the other. It could be a serious mistake, for example, to think one is dealing with critically evaluated data when the data was actually taken from the primary literature with no evaluation attempted.

These secondary sources span the spectrum from the very broad range sources such as International Critical Tables, Beilstein's Handbook of Organic Chemistry or Landolt-Bornstein to more specialized sources which cover narrow areas. Typical examples of specialized sources in the vapor pressure and phase equilibria areas are given as references 2-11 and 12-21, respectively.

THEORETICAL METHODS

In discussing theoretical/calculational efforts, it should be recognized that there are many activities which may fall under this category. This section will be of a general nature with a more detailed description later in the presentation.

The use of these methods are frequently necessary as an addition to or a replacement for experimental data. Even when good experimental data are available it is frequently necessary to represent the data by an equation for ease of interpolation, extrapolation, or as input to various process simulation programs. Frequently this is accomplished by fitting the data to an equation using a least-squares procedure to determine the values for the constants. Other methods include calculation of one property from another using known thermodynamic relationships, e.g. enthalpy from heat capacity or heat of vaporization from the slope of the vapor pressure curve. Another important area is the treatment of experimental binary vapor liquid equilibrium data to obtain the constants for one of the expressions representing excess Gibbs energy, e.g. UNIQUAC, NRTL, Wilson, etc.

Methods for estimation of physical property data have been available for a long time and are continually being improved and expanded. An excellent summary and comparison of many of these methods is given by Reid, Prausnitz and Sherwood (22). Estimation methods are important in cases where it is difficult or impossible to experimentally measure the data of interest and also in cases where a data value is needed quickly for a preliminary calculation with the intent of determining an experimental value to be used at a later date for a more refined calculation. Examples of types of properties which are frequently estimated as a result of difficulty in making accurate measurements are critical parameters and ideal gas heat capacities.

Estimation methods are of various types including group contribution, corresponding states, and homologous series methods. One of the exciting advances in recent years has been the advent of group contribution methods such as UNIFAC (23) and ASOG (24) for the estimation of liquid phase activity coefficients. Although they must be used with care, these methods can achieve high accuracy and they make it possible to predict the vapor-liquid or liquid-liquid equilibria behavior of multicomponent systems from a knowledge of vapor pressure and molecular structure of the pure components. The UNIFAC method has also been applied to calculation of mixture flash points (25) and evaporation rates (26).

EXPERIMENTAL METHODS

The capability of making experimental measurements on systems of interest for process design is of critical importance for the data design laboratory. There is frequently no substitute for direct experimental measurements on the system in question. Of course, it is desirable to have methods readily available with the following attributes: widely applicable, rapid to use, accurate, wide capability in temperature and pressure. As mentioned in the introduction, it is wise to consider setting up apparatus capable of handling many different types of systems. However, there will always be the occasional system which will be inaccessible because of temperature or pressure limitations or corrosivity or toxicity of the chemicals involved. For these cases, one can build specialized apparatus or go to external laboratories which may have the required experimental capability available; or rely on theoretical methods.

Some of the methods available in our laboratory are outlined in Table 2. The method of choice for vapor pressure is usually the twin ebulliometer in the pressure range 10-800 mmHg due to its high accuracy. Other methods available in our laboratory include: vapor-liquid equilibria by an isobaric twin ebulliometric method using an Ellis still or by a semi-automated isothermal static method; cooling curve analysis for sample purity, freezing point, and phase diagram determinations; precision rotating bomb oxygen combustion calorimetry for heat of combustion/formation determinations; high temperature-high pressure flammability measurements in a 36ℓ steel vessel; reaction calorimetry in a number of devices including one capable of operating at the high temperatures and pressures frequently encountered in process conditions. The key to the accuracy of many of these methods is the precise measurement of temperature using precision platinum resistance thermometry (27) or quartz thermometers. In addition, for nearly all experimental capability, our laboratory has a corresponding theoretical capability.

CRITICAL EVALUATION

Critical evaluation is the process of determining the accuracy and reliability of the data and selecting the "best" values. This is essential whether the data was taken from the literature, was determined experimentally for the application of interest, or was estimated. This step is necessary to select the most reliable data from a data set which may contain values of widely varying reliability. Critical evaluation is also necessary even if only a single data set is available since its reliability must also be evaluated. Details of the evaluation process would not be appropriate here but major items to be considered include sample purity, details for calibration of measurement apparatus, general accuracy of method used versus accuracy of other methods, internal consistency of data from each study, agreement with data by other investigators using different methods, and reputation of the investigator for making measurements of high accuracy.

Two examples of critical evaluation activities based on the activities of our group will be briefly discussed. Each makes extensive use of theoretical concepts. The first of these is the JANAF Thermochemical Tables project (28) which is an ongoing project and has been done by our group under government contract since its inception in 1959. This project involves a critical evaluation of literature thermodynamic data along with the requisite calculations to produce a self-consistent set of

thermochemical tables. Originally intended to serve as a data base for military rocket performance calculations, these tables now serve many other applications in process feasibility and analaysis in the chemical industry and in energy technologies. While the JANAF tables are primarily limited to inorganic species, our group has constructed a large number of organic thermochemical tables as required for internal company needs. Many of these organic tables were published by Stull, Westrum and Sinke (29).

The second example is more directly tied to process engineering needs. This involves the establishment of critically evaluated pure compound physical property data bases. Several of these are available commercially for purchase or lease including those from Engineering Science Data Unit (ESDU), Thermodynamics Research Laboratory (Washington University), Thermodynamics Research Center (Texas A&M University), and the one under development by AIChE as part of the Design Institute for Physical Property Data (DIPPR). In addition, each company may have a proprietary data base which is based on a combination of literature and proprietary data.

At Dow, a centralized proprietary computerized Physical Property Data Bank (PPDB) has existed for a number of years and now contains data for in excess of 900 pure compounds. The selection, correlation and entry of data is restricted to two groups of personnel with experience in those operations but access to retrieve the data is open to all employees. Retrieval is through a number of computer programs designed to serve different purposes. One program is for retrieving data in specified units over desired temperature or pressure ranges. Quality codes and temperature ranges are designed to provide the user some indication for the reliability of the data and whether it is derived from experiment or theory. The other programs which access the data base include programs for distillation column design, the Dow proprietary process simulation program (DOWSIM) and the ASPEN PLUS simulation program. All data are stored in PPDB as coefficients for temperature-dependent equations with emphasis on reduced property equation forms. Coefficients for an equation of state are stored for each compound so that real gas properties may be calculated at any desired temperature and pressure.

TYPES OF THEORETICAL EFFORTS

The data design laboratory can serve a unique role in the chemical industry in terms of its ability to supply high quality thermophysical data for a number of uses. Its value is greatest when there are no gaps in the data base. This can be assured by making good use of theoretical techniques as well as experimental information. All data stored in this data base has a quality code associated with it, regardless of the origin of the data. Thus an engineer, in easily accessing and using the data, only needs to ensure that the uncertainties of the data are acceptable within the confines of the current project.

As mentioned earlier, theoretical efforts are often not viewed with the respect they deserve. Such activity is used more often than its prime detractors think. In general terms, such activity can be used to supply missing data, to analyze and critically evaluate single and multiple data sets, and to derive project results from the available data. These types of activity are often called by different names but in fact are all based on good theoretical foundations. [Of course, some foundations are firmer than others.] Four examples will help illustrate the extensive use and the extensive need for theoretical endeavors -- the equation of state of o-dichlorobenzene, the heat capacity of calcium, the analysis of the boron vapor pressure data, and the thermochemical table for gaseous diatomic boron.

EQUATION OF STATE OF O-DICHLROBENZENE

The physical properties of o-dichlorobenzene as a function of temperature and pressure are derived (in our data base) from an approximate equation of state. From this equation of state the normal thermodynamic relationships are used to derive temperature- and pressure-dependent properties such as heat capacity, vapor pressure, and others. The critical properties of o-dichlorobenzene are needed in mathematical description of the equation of state. For o-dichlorobenzene, these properties (critical temperature, critical pressure, and critical volume) have not been measured experimentally; they were estimated. Many estimation techniques were available; the best estimation technique for this type of compound was chosen. These estimated values are combined with other relevant data -- vapor pressure, heat of vaporation, heat capacities, and more -- to arrive at the best compromise of the available data -- measured or calculated -- which will describe properties of o-dichlorobenzene. Since the qualitative behavior of these properties is known as a function of temperature and pressure, the reasonableness of the results can then be verified. In this limited description, there have been five uses of theoretical efforts. They are:

1. determination of the form of an approximate equation of state
2. estimation of critical properties
3. derivation of thermodynamic properties from equation of state
4. simultaneous (statistical) treatment of data -- experimental and theoretical-- to define coefficients in the equation of state and correlating equations
5. verification of final results

Many people would view only the first two items as a theoretical effort. However, all steps involve an interdisciplinary use of data and techniques.

HEAT CAPACITY OF CALCIUM

The heat capacity of calcium is in desperate need of current experimental study. The existing data, except for a study below 10 K, are all based on samples of questionable purity. A thermochemical table can be derived from these data, with heat capacity values in the region between the low temperature heat capacity and high temperature enthalpy measurement being estimated. However, the current best values for calcium use the experimental data as a guide but, in fact, are estimated from the corresponding values for magnesium and strontium.

The basis for the estimation is the fact that the heat capacity curves as a function of temperature for members of the same group in the periodic table of elements are similar. Even though magnesium has a different crystal structure from calcium and strontium, the temperature dependent heat capacity values are expected to exhibit a family trend at least up to 300-400 K. This family trend approach is a very common technique and is a simple and useful method for estimating missing information or verifying the reasonableness of existing data. The approach may be used for many physical/thermodynamic properties.

VAPOR PRESSURE OF BORON

In my previous presentation, the numerous studies of the vapor pressure of boron were discussed. These multiple data sets cover different temperature and pressure ranges (overlapping and non-overlapping data sets) with each study containing a vastly different density of data points. There are many problems associated with the analyses of these multiple data sets. The emphasis in this presentation is to highlight uses of theory to provide quality data, not all the possible pitfalls in any analysis. The most efficient way to evaluate, in a preliminary way, the agreement, or rather the disagreement, of the various studies is to construct a graph of the actual experimental data. The usual representation is the logarithm of the pressure versus the reciprocal of the absolute temperature, log p vs (1/T). In our laboratory this is done by computer; the graphs being roughly 60 x 75 cm. The graphs published in journals (if they are published at all) are often reduced to such a small size that it is difficult to distinguish the data points of the various data sets. The graphs provide an easy visual assessment as to the possible problems in the choice of the real vapor pressure of boron. Such an assessment is not as easily apparent through a comparison of equation coefficients or a table of smoothed values.

The purpose for which this data was collected together will in large part dictate the mode of analysis of these multiple data sets. In any event, there are two theoretical techniques to be used in the analysis -- a mathematical theory (statistics, in this case) and a chemical theory (thermodynamics, in this case).

As a possibility, assume that a reliable description of the vapor pressure and the heat of sublimation of boron is the prime interest. This can be accomplished by using an equation to fit the data. A mathematical procedure will yield coefficients for the equation of interest. Thermodynamic functions provide the relationship between vapor pressure and heat of sublimation. For convenience in meshing the equation with the thermodynamic relationship, an equation of the form $\log p = A + (B/T)$ is appropriate, with the coefficient B being proportional to the heat of sublimation. Proper use of statistics and chemical thermodynamics would permit a selection of reliable vapor pressure and prediction of the heat of sublimation of boron.

Numerous mathematical procedures are available for representing the data by an equation. The primary interest lies in the determination of the coefficients of the equation and their uncertainty. In addition, the degree of confidence in the form of the equation is important. The choice of the specific mathematical procedure should be consistent with the error characteristics of the data. However, a least squares approach is often blindly used. To depict vapor pressure, the equation $\log p = A+(B/T)$ is often used for convenience (as mentioned earlier) but this is not necessarily the best as far as the mathematical analysis is concerned.

A better description of the true vapor pressure of boron may be obtained by the analysis of the experimental data simultaneously with other thermodynamic information. [This latter information could be derived from experimental studies, such as heat capacity measurements, or estimated.] The so-called second and third law combine techniques of statistics and thermodynamics in this evaluation. This procedure provides a better overall consistency between numerous properties for an extended temperature range. The statistics are used for least squares fitting of data and the error analysis whereas thermodynamic relationships are used to interrelate the properties.

GASEOUS DIATOMIC BORON

This example will be used to illustrate the various levels of sophistication which can be used in generating a thermochemical table for gaseous diatomic boron. All of these approaches rest on a good theoretical foundation, but, like experimental data, they have different errors associated with them.

The thermochemical table for $B_2(g)$ is generated by the use of mathematical relationships derived from statistical mechanics and chemical thermodynamics. The data required for these calculations is spectroscopic information -- electronic energy levels and the vibrational-rotational energy levels for each electronic energy. These data may be derived from experiment or theory.

From an extremely simplistic point of view, there are many levels of sophistication available in treating this problem. Each provides a "slightly" different result with a "slightly" different temperature dependent error or uncertainty for each property calculated. Each approach is valid with a firm foundation. Once the user decides how good the thermochemical table must be, then the appropriate theoretical predictions and calculation method can be chosen.

A summary of the varying modes of using spectroscopic information of $B_2(g)$ is given below. They are ordered in degree of increasing sophistication. The data used is given.

1. ground state electronic energy level with its vibrational-rotational constants;
2. ground state electronic energy level with its vibrational-rotational constants; and electronic energy of the excited state at 30572.4 cm^{-1}, using vibrational-rotational constants of ground state;
3. ground state electronic energy level with its vibrational-rotational constants; estimate of energy of two low lying electronic states based on ab initio configuration interaction, use three excited state energies, assuming they all are similar to ground state;
4. use of all four electronic energy levels, each with their own vibrational-rotational structure;
5. use of direct summation technique for all four electronic energy levels.

For each mode of including this spectroscopic information there is a slightly different calculational pathway to use. The various calculations each add a significant improvement to the accuracy of the resulting table. The real key for a reliable tabulation is the inclusion of the two presently unobserved electronic energy levels at low energies. In addition, the knowledge of these levels suggests that similar levels be included (by analogy) for $Al_2(g)$. These two levels change the entropy by 1.46 to 6.67 J mol^{-1} K^{-1} in the 500-600 K range.

CONCLUSIONS

Theoretical works are an important part of any project. This activity can be used to supply missing data, to analyze and critically evaluate single and multiple data sets, and to derive project results from the available data. The interdisciplinary use of data and techniques permit the efficient development of any project. However, this does cause problems in that the scientist may not fully appreciate the strengths and weakness of techniques which are derived from disciplines other than the scientists prime discipline. The proper use of statistics in physical chemistry problems is an example. Physical chemists may not have had as complete a training in statistics as is necessary to properly treat physical chemical data.

TABLE 1

Frequently Needed Process Design Data

Vapor Pressure

Heat of Vaporization

Specific Heat/Enthalpy

Density

Thermal Conductivity

Viscosity

Surface Tension

Phase Equilibria
- Vapor-Liquid Equilibrium
- Solubilities

Heats and kinetics of reaction, polymerization, solution, dilution, etc.

Flammability hazard data

Reaction hazard data

TABLE 2

Experimental Methods in Use at Dow Chemical Thermal Group

			Typical Values		
Type of Measurement	Method	T range, °C	Sample Size	Accuracy	Time Required, Day
vapor pressure	twin ebulliometer	amb to 400	15 ml	<0.1%	1/2
vapor pressure	static manometer	amb to 250	2 ml	0.2%	3
vapor pressure	DSC	-180 to 750	2 mg	1%	1/2
sp. heat/enthalpy	DSC	-180 to 750	2 mg	2%	1/2
thermal cond.	transient hot wire	amb to 250	20 ml	1%	1/2

amb = ambient temperature

REFERENCES

1. J. V. Sengers and M. Klein, "The Technological Importance of Accurate Thermophysical Property Information," NBS Special Publication 590, October, 1980.

2. T. E. Jordon, "Vapor Pressure of Organic Compounds," Interscience, New York, 1954.

3. D. R. Stull, Ind. Eng. Chem., 39, 517 (1947); ibid 39, 540 (1947).

4. S. Ohe, "Computer Aided Data Book of Vapor Pressure," Data Book Publishing Co., Tokyo, 1976.

5. T. Boublik, V. Fried and E. Hala, "The Vapor Pressure of Pure Substances," Elsevier, Amsterdam, 1973.

6. J. A. Riddick and W. B. Bunger, "Organic Solvents: Physical Properties and Methods of Purification," Wiley-Interscience, New York, 1970.

7. J. Timmermans, "Physico-chemical Constants of Pure Organic Compounds," Elsevier, New York, Vol. 1, 1950 and Vol. 2, 1965.

8. R. R. Dreisbach, "Physical Properties of Chemical Compounds," ACS, Washington, DC, Vol. 1, 1955; Vol. 2, 1959; Vol. 3, 1961. These are Advances in Chemistry Series #15, 23 and 29, respectively.

9. Engineering Sciences Data Unit, "Physical Data, Chemical Engineering Series," in 8 volumes, current.

10. K. R. Hall et al., Thermodynamics Research Center Hydrocarbon Project, "Selected Values of Properties of Hydrocarbons and Related Compounds," Texas A&M University, current.

11. K. R. Hall et al., Thermodynamics Research Center Data Project, "Selected Values of Properties of Chemical Compounds," Texas A&M University, current.

12. J. Gmehling, U. Onken et al., "Vapor-Liquid Equilibrium Data Collection," DECHEMA Chemistry Data Series Volume I (in 7 parts).

13. J. M. Sorensen and W. Arlt, "Liquid-Liquid Equilibrium Data Collection," DECHEMA Chemistry Data Series Volume V (in 3 parts).

14. H. Knapp, R. Doring, L. Oellrich, U. Plocker and J. M. Prausnitz, "Vapor-Liquid Equilbria for Mixtures of Low Boiling Substances," DECHEMA Chemistry Data Series Volume VI, 1982.

15. E. Hala, I. Wichterle, J. Polak and T. Boublik, "Vapor-Liquid Equilibrium Data at Normal Pressures," Pergamon, Oxford, 1968.

16. J. J. Christensen, R. W. Hanks and R. M. Izatt, "Handbook of Heats of Mixing," Wiley-Interscience, New York, 1982.

17. I. Wichterle, J. Linek and E. Hala, "Vapor-Liquid Equilibrium Data Bibliography," Elsevier, Amsterdam, 1973; three supplements are available from the same publisher.

18. A. Marzynski et al., "Verified Vapor-Liquid Equilibrium Data," PWN-Polish Scientific Publishers, Warsaw, in 6 volumes.

19. J. Wisniak and A. Tamir, "Liquid-Liquid Equilibrium and Extraction," and Extraction," Vol. 7A, 7B, Elsevier, Amsterdam, 1980; a bibliography.

20. J. Wisniak and A. Tamir, "Mixing and Excess Thermodynamic Properties," Elsevier, Amsterdam, 1978; a bibliography.

21. H. Stephen and T. Stephen, "Solubilities of Inorganic and Organic Compounds," Vol. 1 and 2, Pergamon, Oxford, 1954. Volume 3 of above series was authored by H. L. Silcock and published in 1979.

22. R. C. Reid, J. M. Prausnitz and T. K. Sherwood, "The Properties of Gases and Liquids," 3rd ed., McGraw-Hill, New York, 1977.

23. A. Fredenslund, J. Gmehling and P. Rasmussen, "Vapor-Liquid Equilibria Using UNIFAC," Elsevier, Amsterdam, 1977.

24. K. Kojima and K. Tochigi, "Prediction of Vapor-Liquid Equilibria by the ASOG Method," Elsevier, Amsterdam, 1979.

25. J. Gmehling and P. Rasmussen, Ind. Eng. Chem. Fundam. 21, 186 (1982).

26. A. L. Rocklin and D. C. Bonner, J. Coatings Technol. 52, 27 (1980).

27. D. R. Stull, Ind. Eng. Chem. 18, 234 (1946).

28. D. R. Stull and H. Prophet, "JANAF Thermochemical Tables," 2nd ed., NSRDS-NBS 37, U. S. Government Printing Office 1971; supplements J. Phys. Chem. Ref. Data 3, 311 (1974); ibid 4, 1 (1975); ibid 7, 793 (1978); ibid 11, 695 (1982).

29. D. R. Stull, E. F. Westrum and G. C. Sinke, "The Chemical Thermodynamics of Organic Compounds," Wiley, New York, 1969.

DISSEMINATION OF DATA AND INFORMATION

Dr John R Sutton

Scientific and Technical Information Unit
Department of Trade and Industry
Ebury Bridge House
Ebury Bridge Road London SW1W 8QD

ABSTRACT

The differing needs of several different kinds of data users are considered. These should influence the format and packaging of the data product and the choice of distribution channels. Numerical data is used differently from non-numerical information and the data provider needs to know how his customers use his product. The presentation of original data to assist the compiler and evaluator is considered. Tagging and flagging may make it easier to find data.

1 INTRODUCTION

Data is generated, compiled and evaluated in order to be used. But before it can be used it must be disseminated to those who will use it. It is self-evident that users are not all alike, and their needs will clearly differ with the use they expect to make of data. The effects this has on the most effective methods of packaging and disseminating the data are less obvious, and it is worth considering the way in which different user needs and different types of data may be accommodated.

2 THE NEEDS OF THE INDUSTRIAL DESIGNER

In industry a designer may foresee a need for certain types of data before he begins a design task. The need for other items may only become apparent in the course of the design. The former can be collected over a period of time. The latter will be needed urgently in order not to delay the completion of the task. A company may find it worthwhile to put together collections of the data most likely to be needed by its designers in order to minimise this problem but the urgency generally predisposes designers to favour sources which they anticipate will provide quick and convenient answers.

A survey of 5000 members of the Institution of Mechanical Engineers was reported in 1967 by Woods and Hamilton. They found that engineers relied heavily on personal contacts as sources of information and when using written sources designers looked first to handbooks and textbooks, then to scientific and technical journals and relatively little to other sources. In 1969 the then Ministry of Technology in the UK surveyed the information requirements of engineering designers. The most used sources were the designers colleagues, data from suppliers, handbooks, standards, previous designs and company data collections. As suggested most of the sources were chosen because the designer expected from previous experience that that source would prove helpful. At the time of these surveys computer data banks were little used. Both surveys showed preferences for compilations of data prepared with the designers needs and ways of thinking in mind.

More recent studies confirm these preferences. Chakrabarti, Feineman and Fuentevilla studied the sources of scientific and technical information used by several hundred staff in a large US company. They found the most used source of information was, again, people in the work group closely followed by handbooks, trade journals and newspapers. Experts in the firm, technical reports, reference books and journals were also used frequently. The frequency of use was correlated with availability, ease of use, and expected utility.

Nowadays the design process is likely to involve calculations carried out by computer, and data is required as input to these programs. The calculation cannot proceed until some numerical value is available for each data item. If the data can be obtained in machine readable form, with the items in the correct format and order, time can be saved and errors avoided. If not the designer must key in all of the data from written sources. The mathematical nature of the calculation will determine whether particular

items of data are critical for a satisfactory result. The consequences of wrong results will usually not be symmetrical. (Too small a plant will not do the required job, too large a one will waste money on capital and running costs.) So the designer will look for data collections with sufficient accuracy in his area of interest. New proposals to compile materials data in this way are discussed elsewhere at this meeting.

A design team will, in fact, work with (at least) two distinct types of data. In process design accurate data on physical and chemical properties and reactions will be sought. On the other hand in the mechanical design of a plant to realise a particular process only code data can usually be used. These are data from materials standards and specifications agreed as indicating either typical values or agreed lower or upper bounds of probable values of properties to be expected from materials of a particular specification. Their use implies that a safe account has been taken of the variability of properties within nominally similar materials. This type of data is, in principle, much easier to find than process design data.

To contrast the needs of different users we may consider a single type of information, data on vapour liquid equilibria (VLE) in a fluid mixture. A distillation column is designed to separate the mixture into two parts, one containing most of one or more volatile components, the other containing mostly the other components. The VLE data contain information on the degree of difficulty of this separation and are used to calculate the size of the column needed to separate the mixture at the planned flow rate.

3 THE NEEDS OF THE RESEARCH WORKER

Turning now to the research user, he will have developed a deep knowledge of some highly specific topic - for example the theory of intermolecular forces and the experimental measurements from which they can be calculated. He will know, or know of, the other researchers on his topic and their publications. In the development of a new theory or experiment he will discover a need for information or data in other areas - mathematical and computerised methods, temperature or pressure control and measurement etc. As a professional researcher he is well able to extract what he needs from compilations and review or from the original publications. He will wish to compare the results or opinions of more than one source to arrive at a 'best available' value, and to judge the probable reliability of the data.

A survey of members of the UK professional societies of physicists and chemists, mostly working in research and development, found that they relied mainly on written sources especially original papers, reviews and abstracts. This contrasts with the designers preference for personal contacts, handbooks etc.

In the VLE data example, measurements at a number of compositions define two curves, one for the vapour boundary and the other for the liquid boundary. Mathematical curve fitting may be used to fit the curves to the available data from several different original papers. However the best possible fit from this point of view which will be sought by the research worker, may not be the fit which best answers the designers need to know the degree of difficulty of the separation. A data compiler with a research background may look at the data from a different viewpoint from the designer unless he takes the trouble to consider the designer's needs and methods of working.

4 THE NEEDS OF THE ADMINISTRATOR

A research administrator, perhaps the secretary of a committee taking decisions on the allocation of research resources, needs information on what has already been done, what is being done by others, what is known about the needs in that area of research. This is not primary data, rather data about data and its acquisition may present problems of its own.

In the VLE data example, a proposal to measure the VLE data for a number of mixtures raises the question whether these particular mixtures have been measured before and, if so, what is wrong with the previous measurements. They may be over a different or more limited range of temperatures and pressures, or they may be known or believed to be inaccurate perhaps because there are evident discrepancies. The question also arises whether designers or other users need the data, or the range or accuracy of data proposed, in order to support their decisions.

A planner or policy maker also needs to consider a wide variety of data. Some of these will be specifically related to planning - what has been done in the past, what do similar or competing organisations do, what is the position in other countries? Others

will relate to legal or environmental constraints, to assessments of feasibility or probability of technical success and to the probable financial consequences of the planned action. Allowance may be made for several different assumptions about the national economy, the activities of competitors, suppliers and customers etc. The planner is an example of a multidisciplinary data user but the outcome of his activities may often seem to be little influenced by the type of data which we are primarily considering.

5 REGULATION AND HAZARD DATA

The technical staff of a regulatory agency need to assess the degree of risk of pollution or some other hazard to those working on an industrial process, to transporters or users of the product or to the general public. This might involve data on levels of toxicity determined by experiments with animals or cells, data on volatility or flash-point etc. Here it will be important to know of the most recent data and to know which data has special status in relation to laws or regulations. Much of the data will be supplied by companies subject to regulation, but data from other sources will be needed for comparison and to assist decisions on priorities, what data to require from companies, what supporting research programmes should be pursued, etc.

In industry there arises a corresponding need to know what the regulations require, what data and information is needed to meet regulatory requirements, how much of this is already available and what additional tests are necessary.

This is another multidisciplinary area, and again the nature of the data may sometimes be rather special. For example, a current regulation which states that a particular material or process or procedure is hazardous in a particular manner, or allowable under particular conditions, provides the ruling data on the subject, whether or not other perhaps more recent or more accurate sources suggest that a revised opinion might be possible. Until there is agreement to amend the regulation, the other data cannot be used.

Those with the responsibility for dealing with major incidents for example fires, explosions or spillages involving dangerous chemicals need data relating to the properties and safe handling of such substances. The particular need here is for clear and immediate presentation of the essential information in a suitable form for use in the emergency. If the data is held at a central point good communications will be essential as will 24 hour availability. Local databanks in the control rooms of the emergency services may be preferred and in any event the staff supplying the information should have experience in the fire or other emergency services. There is a multi-disciplinary element in this type of use but the dominating factor is urgency and clarity of communication.

6 STANDARD DATA

Standard data can have several forms, but in each case the intention is to facilitate some aspect of manufacture or commerce by the adoption of agreed values of data. One such application is the use of standardized data to describe the properties of engineering materials mentioned earlier. Standard design codes provide a similar rationalization for the design of heat exchangers, pressure vessels and similar plant facilitating the discussion of structural criteria and promoting safe design.

In other cases the adoption of standard physical properties may be of more direct commercial benefit. For example the transfer of a bulk chemical such as ethylene by pipeline from suppliers to users depends on a measurement of the quantity transferred. An instrument may measure the volume transferred but the mass transfer will depend on the density of the gas at the temperature and pressure occurring at the transfer point and at the time of measurement. It is possible to measure temperature and pressure and consult tables, charts or equations, to determine the density. If supplier and user consult different data they may disagree as to the quantity transferred. The use of agreed tables, charts or equations will avoid this. (However if the agreed data is known to be inaccurate it may still be possible for either supplier or user to manipulate the transfer conditions to their advantage.)

Those involved in determining standards need data which is accurate, comprehensive, up to date and can be made available to all users. There will usually also be a long period of consultation necessary so that the standard can be accepted by users.

7 EXPERT SYSTEMS

A rapidly developing area of artificial intelligence is the construction of expert systems. These contain a store of information about a particular area of expertise eg medical topics, the interpretation of mass spectra, analysis of oil or gas well logs or prediction of the location of mineral deposits. This is usually expressed in the form of rules and referred to as the knowledge base. The expert system also contains programs which develop inferences from the input information by successive application of the rules. It is possible to express complex areas of knowledge in this way by using large numbers of quite simple rules. The process of translating the expert's knowledge into the rules of the expert system is however rather difficult. It has been suggested that 'knowledge engineers' will need computer programming skills combined with the talent of a psychologist to codify the heuristic knowledge of the expert. They will also need to supplement the expert's knowledge with data from conventional sources. The knowledge bases of expert systems are data bases of a special format and as such another type of data usage although detailed discussion is not appropriate here.

8 DISTRIBUTION CHANNELS

To meet the varying needs of these, and many other, users the providers of data need to consider carefully the segments of the market at which their product should be aimed. Who are the expected users and what aspect of their particular needs will be satisfied? What distribution channel will reach the target users? Some will publish their critical appraisal of the measured values of a few related properties for a group of substances in the same scientific journal which carried many of the original data. This should suffice to reach the research user and an awareness of the paper will spread gradually amongst industrial users. Others will offer their work to a specialised data journal whose readers have a particular interest in good data and its thorough evaluation. Other groups might produce reports quickly for distribution by mailing list to industrial users and others who have expressed prior interest, or synchronise their work with the revision cycle of some widely respected handbook.

Professional societies or special interest groups can distribute to their members information and data of particular relevance to them, either compiled for that specific purpose, or independently offered for publication by this channel. They also provide a valuable means of alerting members to the availability of other sources of data which are appropriate and 'friendly' to the particular audience, perhaps new standards, trade literature, handbooks etc. In view of the comments made above on the needs of engineers and designers, this type of route seems to be potentially of great value in promoting awareness of data sources.

Electronic databases with on-line retrieval offer currency of data as one of their advantages and their growing use increases the demand for regular updating. At the same time however they provide a ready and convenient vehicle for the distribution of new data as a matter of course to all who interrogate the data base.

9 FORMAT

As well as the means of distribution of data, it is important to consider the format and presentational style which will be most appropriate for the target user. Tables of numbers are probably the most familiar format for numerical data, and clearly data of any accuracy level from rough guideline to very precise (as in the fundamental constants of physics) can be presented. For data which depend on some independent variable (temperature, time etc) the interval of the variable can be chosen to present the data in any desired amount of detail, and the use of the table is greatly simplified if linear interpolation is sufficient for the level of accuracy needed by the user. Data which depend on several independent variables can be represented by tables but less conveniently, and care is required in the layout, especially if there are more than two independent variables.

Graphs or charts provide a good overall impression of the way in which one variable depends on another. Values can be read off the graph - within accuracy limits dependent on the scale of the graph, care in production etc - for any value of the independent variable within the range of the graph. Several quantities can be represented on the same graph, provided the curves are sufficiently dissimilar not to superimpose and can effectively convey relationships between the two quantities. For example latent heat, entropy difference and volume or density difference between vapour and liquid phases all reach zero at the same temperature and the curves all have vertical

tangents at this point. More complex charts, such as the Mollier and related diagrams of thermodynamic properties, can represent the interrelationships of several quantities in a manner which facilitates a number of different calculations.

For data which depend on location, maps may show the distribution more clearly than would a table. If there are sufficient data and the distribution is believed to be continuous, contours may be calculated and displayed. Colour can greatly enhance the information content of a map. In the case of geodata derived from satellite observations the maps - in digital form - are the original data.

Equations allow the user to calculate a property for any desired value(s) of any number of independent variables - within the range of validity of the equation. The form of an equation may be purely empirical or it may be influenced by theory. When there are interrelations between properties, the equations describing them ought to be mutually consistent. For example equations for latent heat, volume or density and vapour pressure ought to be consistent with the Clapeyron equation. Polynomials can be fitted and evaluated efficiently but are not always suitable - for example no finite polynomial can accurately represent the vertical tangent of the latent heat curve. However some transformation of the variables can usually be found which allows the polynomial form to be used. Equations which rely strongly on exponential functions require caution in extrapolation beyond the range to which they have been fitted. In equations containing several terms only those which make a significant contribution should be included. Statistical tests can, and ideally should, be used to demonstrate the reliability of the proposed equation. This point is discussed in a report by the CODATA Task Group on Data for the Chemical Industry.

Computer algorithms realise an equation in a form suitable for computer calculation. Care needs to be taken to maintain a suitable level of accuracy over the whole range of the independent variable(s) and it is desirable that the calculation should not use up significantly more time or storage space than is necessary. Erroneous input values should always lead to some clear indication of error. The widespread use of micro computers seems to encourage poor programming and may lead to a repetition of many of the mistakes made by the early programmers on the computers of the 1950's and 1960's.

The combination of graphs or charts with algorithms gives the possibility of presenting and manipulating data depending on many variables quickly and easily in a manner which was not possible before computing power was readily available. This has significantly reduced the amount of data which must be determined in order to define a complex system, and a number of such formalisms have been developed in the materials field.

Part of the increased use of computers to handle data, of course, is the growing use of video screens to display data to the user. These may limit the manner and extent to which the data can be displayed especially in the case of videotext. Data which is to be transmitted to the user by videotext frames faces a severe constraint in the frame size, which corresponds to only a fraction of the page size in a book or report. It is not satisfactory simply to divide up a conventional page; the users requirements need to be thought through in sequence and the data format developed for this specific application.

On the other hand the use of computers in database handling can bring significant advantages as far as format (as opposed to display) is concerned. The retrieval software can be used to search the file and display results in many more ways than could be achieved in hard copy. For example, a single file of materials data could be used to display complete collections of data for a specific alloy, or the values of particular properties for a whole range of alloys. It could be searched to provide a particular number relating to an alloy property, or to select materials with property parameters within defined limits. To achieve this flexibility in hard copy would need many sets of tables rather than a single file, although it must be said that with skill and experience manual searching of printed data can achieve remarkable results.

10 COLLECTIONS OR PACKAGES

The producer, the compiler and the validater of data work naturally with related data sets - data on a group of properties which can be handled in similar ways and for a group of substances which present similar problems. The user, especially the

multidisciplinary user, needs to have data which are widely dispersed in original brought together in a convenient collection or package. One way to do this is in large all embrancing collections which seek to be all things to all users. However, the cost of such collections is becoming prohibitive, in addition to the disadvantage of their size, and a better way may be specialist packages tailored to the needs of particular groups of users.

11 PRESENTATION OF ORIGINAL DATA

We have been considering the distribution to the end users of data which have been compiled, evaluated, packaged etc. Between the generation or measurement of the data and the compiling and validation stages there is usually another dissemination stage. The original data is made available in a published paper or laboratory report. The compilers and evaluators need to find and use these papers and reports. Care in presentation by the originator can greatly assist the compiler and evaluator. A CODATA task group on Publication of Data in the Primary Literature with financial support from Unesco prepared a Unesco-Unisist Guide for the presentation of numerical data in the primary literature. The guide was reprinted as Codata Bulletin No 9 and elsewhere. Its recommendations cover the provision of an adequate description of the experimental procedures, so that evaluators can form a view on the quality of the data, and an explanation of the reduction of the instrument readings to be reported results (including any assumptions made and models used), as well as the presentation of the numerical results. Under this last heading the guide recommends that all important numerical results be tabulated in a form as close as practicable to the original measurements, that if smoothed data are also included these should be presented either by equations or by tables with sufficient digits and close enough spacing of the argument to allow interpolation without serious loss of accuracy, and that both the imprecision and the inaccuracy should be presented carefully, distinguishing and estimating the contribution of the various sources. The use of standard symbols, units and nomenclature is also recommended. These recommendations have been interpreted and extended in a series of specialised guides for chemical kinetics, thermodynamics, biochemical equilibrium, biology and a number of other fields. This reinforces the important feature of traceability. While changes in format, packaging, evaluation, etc can all contribute to the ease of identification and use of the data, the means must be preserved through each successive stage of processing for the user who needs to do so to identify the quality of the data he is using, its precision and accuracy and to get back to the original source if necessary. Failure to do this can reduce the value of the data sources: it is misleading at best; at worst it could be positively dangerous.

12 TAGGING AND FLAGGING

As has been stated, compilers and evaluators need to find the original data on which they work. The usual tools of information retrieval, bibliographies, abstracts etc whether in printed form or as computer databases, do not always distinguish between papers which contain new data and those which do not. 'Flagging' means adding a symbol to the abstract of a published paper to indicate that the paper contains data. The symbol may also indicate the broad type(s) of data reported in the paper. 'Tagging' refers to more detailed identification of the data content of a paper. A joint working group of the Committee on Data for Science and Technology and the International Council of Scientific Unions Abstracting Board looked into tagging and flagging in 1976 and expressed the view 'that flagging and/or tagging of the data content of literature is inevitable in the near future'.

The intention of course is to save the reader's time by allowing the retrieval from bibliographic, abstracting and index services of only those papers which do contain data. To be fully effective a scheme of flagging or tagging needs to be a universal one, compatible across scientific disciplines, language independent and well judge in the level of detail. Up to a point more detail, eg the type of material, its state of aggregation, the property measured, the method of measurement etc increases the usefulness of the tag but it also increases the effort required and the cost of providing the tag and of copying it into all the relevant entries in an index or databank. Despite the optimism of the Joint Working Group, progress towards a widely acceptable scheme seems to be very slow. In the meantime authors of papers containing data can help by ensuring that their papers contain useful abstracts in which this is made clear.

13 CONCLUSIONS

There are many ways of disseminating data, some less obvious than others to the primary data producers. By considering the needs and data-seeking habits of several types of

user, we have seen that both the means of dissemination and the format of different types of data should be matched to the target user if the most effective use of the data is to be ensured. Data is both a costly and a valuable commodity and this care in dissemination can improve the benefit-to-cost ratio. Throughout the packaging and dissemination processes the data should always retain sufficient identity to give the user an idea of its quality and the means to go back to the source if required.

REFERENCES

D N Woods and D R L Hamilton 'The Information Requirements of Mechanical Engineers' Library Association 1967.

'Survey of information needs of physicists and chemists' Report of a survey in 1963-64 in association with Professor B H Flowers on behalf of Advisory Council on Scientific Policy. J Documentation 21 (2) 1965 83-112.

J R Sutton 'Information requirements of engineering designers' in AGARD Conference Proceedings No 179 1976.

A K Chakrabarti, S Feineman and W Fuentevilla 'Characteristics of Sources, Channels and Contents for Scientific and Technical Information Systems in Industrial R & D' IEEE Transactions on Engineering Management 1983 EM30 (2) 83-88.

CODATA Bulletin 30 December 1978 'Guide for the presentation in the primary literature of physical property correlations and estimation procedures'.

CODATA Bulletin 9 December 1973 'Guide for the presentation in the primary literature of numerical data derived from experiments'.

CODATA Bulletin 13 December 1974 'The presentation of chemical kinetics data in the primary literature'.

CODATA Bulletin 20 September 1976 'Recommendations for measurement and presentation of biochemical equilibrium data'.
CODATA Bulletin 25 November 1977 'Biologists guide for the presentation of numerical data in the primary literature'.

CODATA Bulletin 32 August 1979 'Guide for the presentation in the primary literature of numerical data derived from observations in the geosciences.

CODATA Bulletin 44 August 1981 'Calorimetric measurements on cellular systems. Recommendations for measurement and presentation of results'.

CODATA Bulletin 46 April 1982 'Guide to the presentation of astronomical data'.

CODATA Bulletin 19 June 1976 'Flagging and tagging data to indicate its presence and facilitate its retrieval.'

ACKNOWLEDGEMENTS

This text draws on the experience of my colleagues in the Department of Trade and Industry to whom I express my thanks.

GENERAL REVIEW OF NUMERICAL DATA BASES

David R. Lide, Jr.
Office of Standard Reference Data
National Bureau of Standards
Washington, D.C. 20234

SUMMARY

Quantitative data resulting from measurement or calculation play a key role in every field of science and technology. The organization of such data into compilations or data bases has become more difficult as the amount and complexity of information in the technical literature increases. This paper surveys current efforts to prepare data bases of interest to the scientific and engineering communities. Emphasis is placed on computer-searchable data bases and their dissemination through on-line networks and other means. Mechanisms for international cooperation are discussed.

INTRODUCTION

The "information explosion" in science and technology has been a popular theme for writers and speakers over the past 25 years. The rapid growth--at times literally exponential--in published papers, reports, and computerized data files has taxed both our institutional and human ability to store, retrieve, and assimilate information. Fortunately, the growing power and availability of digital computers has prevented a complete swamping of our traditional means for handling scientific information. Impressive progress has been made in the last decade in developing computer-based bibliographic files which can be searched interactively from remote terminals. These files permit rapid search of hundreds of thousands of documents on the basis of a profile of key words--a carefully selected set of terms which attempts to match the interest of the user with the contents of the documents. The result of such a search is a list of potentially relevant documents which the user must then acquire and examine. These "on-line" bibliographic search systems, though still not fully perfected, have proved their utility in making a first pass through the technical literature.

What the scientist or engineer ultimately wants, however, is factual information. In most cases, this means numerical data that describe, in quantitative terms, the behavior or properties of some material or system. There is great interest at the present time in extending computer-based information-handling techniques to include such numerical data. This offers, in principle, a way to transmit needed data to the ultimate user by electronic means, without the intermediate step of referring to a book, report, or journal.

CLASSES OF DATA

It is unfortunate, and somewhat ironic, that the scientific and technical information community has failed to develop a precise vocabulary for many key concepts. The term "data base" is frequently used for the bibliographic files mentioned above (Chemical Abstracts, Engineering Index, etc), as well as for numerical data files, although certain groups have attempted to differentiate by calling the latter "data banks." For the purpose of this paper, data will be understood to mean factual information, usually expressed in numerical form, which has been derived from some measurement, observation, or calculation. A data base is an organized collection of such factual information on a well-defined topic. We shall be concerned particularly with data bases expressed in some computer-readable medium.

A detailed classification scheme for scientific and technical data has been presented by the Committee on Data for Science and Technology (CODATA) of the International Council of Scientific Unions (ICSU) through its Task Group on Accessibility and Dissemination of Data[1]. The fine structure of this scheme need not concern us here, but it is helpful to recognize three broad classes of data:

Class A--Repeatable measurements on well-defined systems. This includes the traditional data of physics and chemistry, resulting from measurements of well-understood properties of systems of known composition. In principle, such data are subject to verification by repeating the measurements in a different laboratory at a different time.

Class B--Observational data. Here we include the results of measurements which are time- and space-dependent and cannot, in general, be checked by subsequent remeasurement. This category includes many of the data from the geosciences as well as environmental monitoring data.

Class C--Statistical data. This class includes various data which are not always thought of as "scientific," but which are important in many technical problems--demographic data, records of production of chemicals, energy consumption figures, health statistics, and the like.

The three classes have sometimes been referred to as hard, semi-hard, and soft data, respectively. Clearly, the boundaries between the classes are not sharp. For example, most biological data are time- or space-dependent, but a carefully designed experiment with an adequate sample of organisms can yield results with high enough statistical significance to be placed in Class A.

THE GROWING NEED FOR GOOD DATA

In more placid times than we live in today, the typical scientist or engineer encountered no major problem in locating whatever data existed that were pertinent to his needs. A small number of journals and handbooks, plus the informal peer-group information chain, were generally sufficient. Judgments on the validity of the data so located could often be made on the basis of direct personal knowledge of the laboratory which generated the data. This process becomes less and less tenable as the scale of needs and data sources increases. Equally important, the most crucial and challenging problems that we deal with today usually require data which cut across traditional disciplinary lines. The peer-group channels and personalized value judgments are no longer as effective as they once were.

As a typical example, consider the debates on depletion of the ozone layer by supersonic transport exhaust and release of chlorofluoromethanes. Important political and economic decisions rest on complex scientific questions. Elaborate mathematical models have been developed in an effort to predict, with sufficient reliability, the consequences of certain existing or proposed practices. These models require various types of input data. Meteorological data (Class B) enter into the prediction of transport of the pollutants into the stratosphere. Rate constants for several hundred chemical reactions and cross sections for photodissociation by solar ultraviolet radiation (Class A) determine the complex chemistry occurring in the stratosphere. Long-term data on the ambient ozone concentration as a function of latitude (Class B) are necessary to provide a base-line against which man-made perturbations can be compared. Epidemiological data (Class C) are needed to predict the level of risk to human beings from diseases such as skin cancer induced by higher ultraviolet radiation levels. The numerous studies of the ozone depletion question[2] have rested heavily on these and other types of data.

Other examples are easily found. The proposal to increase our dependence on coal as an energy source raised many questions which can only be answered with adequate data bases. The needs include better data on the physical, chemical, and toxicological properties of compounds derived from coal; base-line data on pollutants likely to be released by expanded coal utilization; and a long list of data relevant to the effects of increasing the atmospheric carbon dioxide level. Data on fatalities from grade-crossing accidents involving coal-carrying trains have even been used in one recent study of energy options[3]. A comparable variety of data needs enters into the consideration of nuclear, solar, and other energy sources.

It is safe to predict that examples such as these will multiply. Bitter debates on the risks from allegedly carcinogenic food additives and contaminants have already occurred. The implementation of the Toxic Substances Control Act of 1976 raised a long series of data needs--not only data on the toxic effects of chemical compounds, but also physical and chemical properties, production and use statistics, and a variety of environmental monitoring data. These needs were explicitly recognized by Congress in Section 25(b) of the act which required the Council on Environmental Quality to coordinate a study of "the feasibility of establishing a standard classification system for chemical substances and related substances and a standard means for storing and for obtaining rapid access to information respecting such substances."[4]

Even aside from questions of health and safety which are subject to government regulation, most industries are dependent on good data bases for process selection, design, and control. Sudden shifts in the cost and availability of raw materials dictate changes in production conditions. The high cost of energy encourages "fine-tuning" of processes to minimize energy consumption and optimize product yield. Long-term reliability of manufactured products becomes a factor in meeting competition. Engineering design is done more systematically, with growing use of computer-based models for simulation of actual processes. The answers obtained from these models are no better than the data that are put in.

THE ELECTRONIC REVOLUTION IN DATA DISSEMINATION

Handbooks, journals, and other traditional publication formats still serve as the major source of data for most scientists and engineers. However, the increasing cost of composition and printing, as well as the difficulty in updating massive hard-copy volumes, are strong driving forces toward the use of modern computer and telecommunications technology for data dissemination. Computer-based formats offer several advantages:

a) It is easier to maintain currency of a data base through frequent updating.

b) More sophisticated search strategies are possible; for example, one may carry out multiparameter searches using Boolean operations which are not practical with printed tables.

c) The data resulting from a search can, with appropriate software and hardware, be put into a computational program for further manipulation without the need for human transcription.

d) Current projections indicate a decrease in storage and telecommunication costs, while all costs associated with printed books are likely to continue to rise.

Computerized data dissemination appears to be developing along two parallel paths: distribution of data bases in some tangible storage medium (e.g., magnetic tapes) and distribution through on-line interactive networks. The former is widely used for certain types of data bases. The World Data Centers[5] provide many items in magnetic tape format, such as oceanographic data and geophysical data obtained from satellite observations. The National Technical Information Service handles magnetic tape distribution for a number of government agencies.[6] The incorporation of data bases in the internal data processing systems of instruments such as mass spectrometers is a growing practice. These systems permit the results of a measurement on an unknown substance to be matched in real time against a library of reliable data, thus greatly facilitating the identification of unknowns.[7]

The distribution-tape mechanism is most attractive when the user expects a continuing, reasonably heavy need for the particular data, which justifies the effort required to install the data base and necessary software on his own computer. When the need is more sporadic, or when the user is unable to make the investment required to handle the data base in his own institution, access through an on-line network becomes more appealing. This requires only a modest capital investment in a suitable computer terminal plus, in some cases, a small subscription or entry fee. Beyond that, the user pays only for the data that he actually retrieves from the network.

On-line information retrieval has become familiar to many scientists and engineers through use of the bibliographic files in services such as Lockheed DIALOG, SDC ORBIT, Bibliographic Retrieval Services (BRS), the National Library of Medicine (MEDLINE, TOXLINE, etc.), and others. Over 2 million individual searches are estimated to have been made in the last year through services such as these, which provide access to several hundred bibliographic files of interest in science and technology. Interest is growing in extending on-line services from bibliographic files to numerical/factual data bases. There are, in fact, several hundred such data bases already available to the public through on-line services. Although the majority are in the business and economic area (e.g., stock market quotations, currency exchange rates, etc.), a significant amount of hard scientific data can be accessed on-line. The existence of these services demonstrates that there are no serious technological barriers to the growth of on-line scientific and technical data bases.

One on-line data service which has become well-established is the Chemical Information System (CIS), which was initiated by NIH and EPA and now includes participation by several other agencies.[8] The CIS now has over 500 subscribers who make a total of about 25,000 searches per month. Some of the currently active components, such as MSSS (mass spectra) and XTAL (single crystal lattice parameters), contain numerical data bases in the sense defined in this paper. Others contain descriptive information which includes some numerical data; for example, OHMTADS gives a concise narrative statement of the degree of hazard to public health of each compound in the file plus quantitative data, when available, on its toxicity. The NMRLIT and FRSS components are bibliographic in nature. Each component includes appropriate software for retrieving the data or other information contained therein.

The components of the CIS are linked to a central hub known as the Structure and Nomenclature Search System (SANSS). This is a file of close to 200,000 chemical compounds containing the formula, name, synonyms, and complete structure record for each compound. The file can be searched on the basis of structure, substructure, full or partial name, or formula. Once a compound is identified through SANSS, the user is informed which CIS components contain information on the compound, and he can be transferred to the component of interest to him by means of a simple command. The Chemical Abstracts Service Registry Number serves as the unique identifier for a chemical compound which permits the linking of the individual data bases to SANSS.

The CIS illustrates the power of an on-line search system to provide a variety of information in a quick and convenient manner. For example, a user might employ the Mass Spectral Search System or the X-Ray Single Crystal Search System to identify an unknown substance. He can then determine from other CIS components whether the substance has been detected in certain environmental monitoring records, its physical properties and toxicity, and whether it has been cited in recent government regulations. Although some of the components are not yet completely developed, the advantages of the CIS concept are already clear. The user gains essentially instantaneous access to diverse classes of information which would be quite laborious to obtain by conventional methods.

Other on-line systems for disseminating numerical data bases are being developed both in this country and abroad. The Metals Properties Council, in collaboration with several other organizations, is planning an on-line materials property system. The American Society for Metals and the National Bureau of Standards are cooperating on plans for on-line dissemination of alloy phase diagrams. Lawrence Livermore Laboratory has developed an extensive on-line system which emphasizes data related to energy storage and conservation.[9] In France an on-line network called Questel is now in operation, and numerical data bases of scientific interest will be included in the EURONET-DIANE network. All observers foresee a very rapid growth in the next five years.

There are, however, certain problems which must be overcome before the potentialities of these on-line systems are realized. Start-up costs are high, and the economics of such systems are very complex.[10] The most efficient and equitable basis for charging users for service is not always clear. Standardization is a major problem. Even with bibliographic services, a user must learn the language and protocols peculiar to each service.

In dealing with numerical data, the more complex file structure and greater variety of formats accentuate this problem. While much attention is now being devoted to designing "user-friendly" software which will minimize the need to master a complex set of instructions, this goal is not likely to be reached quickly or easily.

Finally, there is the central problem of quality assurance. In conventional publications of reference data it is possible to include expository text and footnotes pointing out the accuracy limits and explaining any caveats on the use of data. Also, citations can be given to the full literature from which the data were derived, so that an interested user can track each recommended value to its source. The more rigid constraints imposed on a computerized data file make it difficult to convey such auxiliary information to the user. It is very important, therefore, that the contents of an online data base be carefully verified and that the user be assured that the values he retrieves are the best available. The National Bureau of Standards is following this policy in all its computerized data bases of physical and chemical properties; in general, the computer version will be backed up by publications which document the choices of data and explain the evaluation process. Unfortunately, such a policy has not been clearly established for other types of data. The practice of putting unverified data into publicly-available systems is a very dangerous one.

DATA CENTERS

In the early 1960's the concept of the "information analysis center" (or IAC) became popular. In generic terms an IAC is an institution for collecting, organizing, evaluating, and disseminating information in a systematic way. Some IAC's deal with a specific class of information; others are focused on a specific problem or application that may require various types of information. An IAC that is concerned primarily with numerical data has come to be called a "data center."

Many data centers that deal with observational (class B) data have been in existence for 25 years or more. Examples are the ICSU World Data Centers for geophysical data[5], the National Oceanographic Data Center, National Space Sciences Data Center, and others. This type of data center serves as a repository for data collected by various organizations in the course of their work. The center receives the data in either computer-readable or hard-copy form (tables, charts, etc.), organizes and indexes them, and makes the results available on request. Although there are some exceptions, most data centers of this type do not carry out extensive analysis or screening of the data.

Data centers in the physical sciences and engineering operate somewhat differently. Here the source of most of the data is the primary research literature (plus some less formal literature such as government agency and contractor reports). The data center tries to locate and acquire, on a continuing basis, all documents that may contain data pertinent to its assigned scope. After organization and in-depth indexing of the documents, this bibliographic file becomes a resource for the center to use in preparing numerical data bases. That process involves:

a) extraction of data from the documents

b) conversion to a common set of units, standard conditions, etc.

c) judgment of the quality of each data set

d) correlation of related data sets and testing against theoretical relationships

e) selection of "best values" on the basis of this evaluation process

f) extrapolation, interpolation, or other forms of prediction of property data that have not been subject to direct experimental measurement

Thus the data center carries out a thorough analysis of all pertinent data with the aim of producing a recommended data set for general scientific and technical use.

The primary output of most such data centers is recommended sets of data for public distribution. Some centers also prepare critical reviews which discuss the reliability of different measurement techniques or which point out needs for new experimental measurements. Some publish annotated bibliographies or data indexes which guide the user to the primary literature sources. Most data centers provide some degree of customized service by responding to public inquiries for specific data points.

In the United States the Office of Standard Reference Data of the National Bureau of Standards has served as the focal point for coordination of data centers in the physical sciences. The resulting complex of centers is known as the National Standard Reference Data System (NSRDS)[11]. Some of the NSRDS centers are located in the technical divisions of NBS, while others are in universities and other institutions. Data centers in the area of nuclear physics are managed by the Department of Energy. A list of NSRDS data centers with addresses and phone numbers is given in Appendix I.

NBS also carries out the dissemination of most of the outputs of these data centers. The Journal of Physical and Chemical Reference Data, published jointly by NBS, the American Institute of Physics, and the American Chemical Society, is a major publication channel. Data on alloys appears in the Bulletin of Alloy Phase Diagrams, published jointly by NBS and the American Society for Metals. Various other professional society, government, and commercial publication channels are also used.

The existing data centers provide a valuable resource which, unfortunately, is not known to all those who could make use of it. CODATA has a project underway to prepare lists of data centers and other data sources in various scientific disciplines. The following chapters of the CODATA Directory of Data Sources for Science and Technology have appeared thus far:

	CODATA Bulletin No.
Crystallography	24
Hydrology	35
Astronomy	36
Zoology	38
Seismology	42
Chemical Kinetics	43
Nuclear and Elementary Particle Physics	48
Atomic and Molecular Spectroscopy	49

Other chapters in preparation deal with chemical thermodynamics, geodesy, geomagnetism, ice and snow, and oceanography. Additionally, further chapters in the biological sciences are planned. These Directory chapters are available from Pergamon Press under the CODATA Bulletin number listed above.

PHYSICAL SCIENCE DATA BASES

As discussed earlier in this paper, there is a strong trend toward dissemination of numerical data by computer-based techniques. Thus, we are beginning to see the appearance of computer-readable data bases as one output mode of the data centers described above. At present most of these data bases are distributed in parallel with hard-copy versions of the same data; this practice is likely to continue for the foreseeable future, because there will continue to be a market for conventional printed compilations.

NBS now offers the following data bases to the public in magnetic tape form:

NBS Chemical Thermodynamics Data Base

NBS/NIH/EPA Mass Spectral Data Base

NBS Crystal Data Identification File

NBS Thermophysical Properties of Hydrocarbon Mixtures

The first three of these data bases contain data on large sets of individual chemical substances (15,000, 40,000, and 60,000, in the order listed). The fourth is different in nature; it contains a small data set plus computational programs that provide specified properties of any desired mixture at any temperature and pressure designated by the user.

Other NSRDS data centers are preparing computer-readable data bases in the following areas:

High-Temperature Thermodynamic Properties (JANAF Tables)

Infrared Spectra

Physical and Thermodynamic Properties of Organic Compounds

Molten Salt Properties

Atomic Spectra and Energy Levels

Chemical Kinetics

Alloy Phase Diagrams

Other groups are also active in creating physical science/engineering data bases[12]. The International Atomic Energy Agency supplies neutron cross section and nuclear structure data bases to its members on a regular basis. Scientific Group Thermodata Europe (SGTE) offers computer access to data of metallurgical interest; this is a joint effort of groups in France, West Germany, and the UK. Data for chemical engineering design applications are provided in computer formats by the PPDS system in the UK, EROICA in Japan, and DECHEMA in Germany. The Cambridge Crystallographic Data Centre distributes an evaluated data base on the structure of organic crystals, which is updated three times per year. A number of other activities leading toward the production of data bases are in progress in the United States, Canada, Japan, and Western Europe.

Finally, mention should be made of discussions of an on-line materials data system that would provide access to mechanical and other properties of structural materials of engineering importance[13]. The concept of such a system is being developed jointly by the NBS Office of Standard Reference Data, the Metals Properties Council (an industry-supported group), and several professional engineering societies. NASA and the Defense Department have also indicated interest. While the planning is still in a very early stage, the basic concept is a distributed system through which organizations could make their own data bases available to a wide audience. The user would have a single contact point which would lead him to the particular data base that can answer his query.

INTERNATIONAL COOPERATION

The cost of creating computerized data bases and developing on-line services to deliver them is not trivial. Systematic planning and cooperation can help reduce these costs. While some countries are planning highly centralized on-line information networks, it seems unlikely that the United States will develop such a comprehensive master plan. The private information industry, the not-for-profit societies, and the Federal Government have their own distinct interests and responsibilities, so that a pluralistic approach is inevitable. There are strong arguments, in fact, that no one can predict the optimum design for on-line data dissemination systems at this time. A reasonable level of competition is desirable in order to provide users with a variety of options and allow the marketplace to decide which are most effective. Nevertheless, some central coordination is desirable, if only to encourage a degree of compatibility which will benefit all services.

While such arguments can be applied to the systems for delivery of the data, it is more difficult to justify duplication in the creation of the data bases. This is a very expensive process, especially if adequate quality control is exercised. Thus, there are strong incentives for collaboration in building data bases and for reaching agreement among interested organizations so as to avoid duplication of effort.

The evolution of on-line bibliographic services may provide a useful model for developing on-line data systems. Several services now coexist in the United States, and others are coming on-line elsewhere in the world. All of these services offer the same bibliographic files in chemistry, physics, biology, engineering, and other major disciplines, but they differ significantly in their search software and other operating features. From the users' viewpoint, this diversity is a great advantage. However, no one would suggest duplication of the enormous effort required to create, for example, the Chemical Abstracts file. On-line data dissemination should follow a similar pattern, with the same numerical data bases available on several different systems, each perhaps emphasizing a different type of application or a different group of users. Services which combine numerical data bases with bibliographic files and other information may also prove attractive. There appear to be many opportunities for creative thinking in the design of on-line services; what is most needed at this stage is rapid development of a sufficient number of good data bases to go into these services.

Cooperation at an international level can also expedite the development of on-line services and help reduce the costs. The ICSU World Data Centers provide a useful precedent for exchange of geophysical data in an organized fashion between distribution points in different countries or regions. The International Atomic Energy Agency has performed a similar function for nuclear and other data relevant to reactor development. In 1966 ICSU established CODATA as a standing organization concerned with various aspects of data collection, evaluation, and dissemination. CODATA has served as a useful umbrella for international efforts to standardize data handling and presentation. Through its biennial conferences, CODATA also brings together scientists from all disciplines who are concerned with data problems. CODATA is expected to play an important role in promoting international cooperation in development of the computer-based dissemination systems of the future.

CONCLUSION

Predicting the changes in data dissemination and use over the next decade requires considerable courage. Nevertheless, certain trends are clear. Among these are:

Needs for reliable data will become more pressing. Many political and economic decisions important to our society will rest on scientific and technical data. Moreover, the pressure for greater productivity and efficiency in industry will create requirements for better data for use in industrial design and process control.

Computer-based data dissemination methods, especially on-line systems, will grow in use. This process will be accelerated by the entrance of younger scientists who have become comfortable with computer terminals during their education. The printed handbook is not likely to disappear but will gradually become a by-product, through the already existing automated typesetting technology, of the same computer-readable data bases which are available on on-line networks.

Coordination in the development of computer-based systems will be essential. There is danger of wasteful duplication in the creation of machine-readable numerical data bases which could greatly increase the overall costs of implementing on-line systems, which will in any case be very expensive. Furthermore, the linkage of related on-line systems will require careful attention to questions of compatibility and format standardization. This coordination must be carried out at an international level.

REFERENCES

1. M. Kotani, Ed., Study of the Problems of Accessibility and Dissemination of Data for Science and Technology, CODATA Bulletin No. 16 (1975).

2. Environmental Impact of Stratospheric Flight, National Academy of Sciences (1975); Stratospheric Ozone Depletion by Halocarbons: Chemistry and Transport, National Academy of Sciences (1979); The Stratosphere: Present and Future, National Aeronautics and Space Administration (1979).

3. Energy in Transition, 1985-2010, National Academy of Sciences (1980).

4. The Feasibility of a Standard Chemical Classification System and a Standard Chemical Substances Information System, July 1978, available from the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402 (Stock No. 041-011-00039-4).

5. Fourth Consolidated Guide to International Data Exchange through the World Data Centers, issued by the Secretariat of the ICSU Panel on World Data Centers, Washington, D.C. (1979); further information may be obtained from the Geophysics Research Board, National Academy of Sciences, 2101 Constitution Av., Washington, D.C. 20418.

6. Directory of Computer Software and Related Reports, NTIS, 1980, National Technical Information Service, 5285 Port Royal Rd., Springfield, Va. 22161.

7. A data base containing mass spectra of over 30,000 compounds is available in magnetic tape form through the Office of Standard Reference Data, National Bureau of Standards, Washington, D.C. 20234.

8. G. W. A. Milne and S. R. Heller, J. Chem. Inf. Comput. Sci. 20, 204 (1980); S. R. Heller and G. W. A. Milne, Database 3, 45 (1980); Environ. Sci. Technol. 13, 798 (1979).

9. V. E. Hampel, in Proceedings of the Seventh International CODATA Conference, P. Glaeser, Ed. (Pergamon Press, Oxford, 1981).

10. S. R. Heller, in Proceedings of the Seventh International CODATA Conference, P. Glaeser, Ed. (Pergamon Press, Oxford, 1981).

11. Further information on the NSRDS may be obtained from the Office of Standard Reference Data, National Bureau of Standards, Washington, D.C. 20234.

12. J. Hilsenrath, Summary of On-Line or Interactive Physico-Chemical Numerical Data Systems, Nat. Bur. Stand. (U.S.), Tech. Note 1122 (1980).

13. J. H. Westbrook and J. R. Rumble, Jr., Eds., Computerized Materials Data Systems (1983), available from the Office of Standard Reference Data, National Bureau of Standards, Washington, D.C. 20234.

APPENDIX I: NSRDS DATA CENTERS

Alloy Phase Diagram Data Center

Dr. Kirit Bhansali
Center for Materials Science
Materials Bldg. - Room B266
National Bureau of Standards
Washington, D.C. 20234
Telephone: (301) 921-2811

Aqueous Electrolyte Data Center

Dr. B. R. Staples
Center for Thermodynamics and Molecular Science
Chemistry Bldg. - Room A164
National Bureau of Standards
Washington, D.C. 20234
Telephone: (301) 921-3632

Atomic Collision Cross Section Information Center

Dr. Jean Gallagher
Joint Institute for Laboratory Astrophysics
University of Colorado
Boulder, Colorado 80309
Telephone: (303) 492-7801

Atomic Energy Levels Data Center

Dr. W. C. Martin
Center for Radiation Research
Physics Bldg. - Room A167
National Bureau of Standards
Washington, D.C. 20234
Telephone: (301) 921-2011

Atomic Transition Probabilities Data Center

Dr. W. L. Wiese
Center for Radiation Research
Physics Bldg. - Room A267
National Bureau of Standards
Washington, D.C. 20234
Telephone: (301) 921-2071

Center for Information and Numerical Data Analysis and Synthesis (CINDAS)

Dr. C. Y. Ho
Purdue University
CINDAS
2595 Yeager Road
West Lafayette, Indiana 47906
Telephone: (317) 494-6300
Direct inquiries to: Mr. W. H. Shafer

Chemical Kinetics Information Center

Dr. R. F. Hampson, Jr.
Center for Thermodynamics and Molecular Science
Chemistry Bldg. - Room A166
National Bureau of Standards
Washington, D.C. 20234
Telephone: (301) 921-2565

Chemical Thermodynamics Data Center

Dr. David Garvin
Center for Thermodynamics and Molecular Science
Chemistry Bldg. - Room A152
National Bureau of Standards
Washington, D.C. 20234
Telephone: (301) 921-2773

Crystal Data Center

Dr. A. D. Mighell
Center for Materials Science
Materials Bldg. - Room A221
National Bureau of Standards
Washington, D.C. 20234
Telephone: (301) 921-2950

Diffusion in Metals Data Center

Dr. John R. Manning
Center for Materials Science
Materials Bldg. - Room A153
National Bureau of Standards
Washington, D.C. 20234
Telephone: (301) 921-3354

Fluid Mixtures Data Center

Mr. N. A. Olien
Center for Chemical Engineering
National Bureau of Standards
Boulder, Colorado 80303
Telephone: (303) 947-3257

Fundamental Constants Data Center

Dr. Barry N. Taylor
Center for Absolute Physical Quantities
Metrology Bldg. - Room B358
National Bureau of Standards
Washington, D.C. 20234
Telephone: (301) 921-2701

Fundamental Particle Data Center

Dr. Robert Kelly
Lawrence Berkeley Laboratory
University of California
Berkeley, California 94720
Telephone: (415) 486-5885

High Pressure Data Center

Dr. Leo Merrill
P.O. Box 7246
University Station
Provo, Utah 84602
Telephone: (801) 378-4442

Ion Energetics Data Center

Dr. Sharon Lias
Center for Thermodynamics and Molecular Science
Chemistry Bldg. - Room A139
National Bureau of Standards
Washington, D.C. 20234
Telephone: (301) 921-2439

Isotopes Project

Dr. Janis Dairiki
Lawrence Berkeley Laboratory
University of California
Berkeley, California 94720
Telephone: (415) 486-6152

JANAF Thermochemical Tables

Dr. Malcolm W. Chase
Dow Chemical Company
1707 Building
Thermal Research Laboratory
Midland, Michigan 48640
Telephone: (517) 636-4160

Molecular Spectra Data Center

Dr. F. J. Lovas
Center for Thermodynamics and Molecular Science
Physics Bldg. - Room B268
National Bureau of Standards
Washington, D.C. 20234
Telephone: (301) 921-2023

Molten Salts Data Center

Dr. G. J. Janz
Rensselaer Polytechnic Institute
Department of Chemistry
Troy, New York 12181
Telephone: (518) 270-6344

National Center for Thermodynamic Data of Minerals

Dr. John L. Haas, Jr.
U.S. Geological Survey
U.S. Department of the Interior
959 National Center
Reston, Virginia 22092
Telephone: (703) 860-6911

Phase Diagrams for Ceramists Data Center

Dr. Lawrence P. Cook
Center for Materials Science
Materials Bldg. - Room A227
National Bureau of Standards
Washington, D.C. 20234
Telephone: (301) 921-2844

Photon and Charged-Particle Data Center

Dr. Martin J. Berger
Center for Radiation Research
Radiation Physics Bldg. - Room C313
National Bureau of Standards
Washington, D.C. 20234
Telephone: (301) 921-2685

Radiation Chemistry Data Center

Dr. Alberta B. Ross
University of Notre Dame
Radiation Laboratory
Notre Dame, Indiana 46556
Telephone: (219) 239-6527
FTS 333-8220

Thermodynamics Research Center

Dr. Kenneth R. Hall
Thermodynamics Research Center
Texas A & M University
College Station, Texas 77843
Telephone: (409) 845-4940

Thermodynamic Research Laboratory

Dr. Buford Smith
Department of Chemical Engineering
Washington University
St. Louis, Missouri 63130
Telephone: (314) 889-6011

Progress Toward a Coordinated System of Databases Covering the Engineering Properties of Materials

J.H. Westbrook
Materials Information Services
General Electric Research and Development Center
Schenectady, NY 12305 USA

Three studies - one by the Metal Properties Council, an ad hoc Materials Data Workshop and a committee report by the National Materials Advisory Board - all show the great need for computer access to engineering properties information on materials. All find such a system to be technically within the capabilities of modern computerized information technology. What is now required is a massive and continuing cooperative effort by a diverse group of stakeholders in such an information system.

INTRODUCTION

Ready access to, and established reliability of, engineering properties of materials are essential to effective materials selection and application, process control, product reliability and performance - all of which contribute to the vigor and growth of industrial economies. Modern information technology, especially the computer, is revolutionizing the whole engineering function from design through manufacturing. Unfortunately, one critical link is missing: computerized files of materials properties data which are needed to fully integrate computer-aided-design (CAD), computer-aided-manufacturing (CAM) and computer-aided-testing (CAT).

In most cases a computer file of materials data is not merely a compact store for a large volume of information and a convenient, versatile means of access, but also an opportunity for enhanced engineering capabilities and a necessity for some design analysis. In addition, the computer potentially offers inherent advantages of comprehensiveness, currency, and speed and accuracy of search that are difficult for a "system" of printed handbooks to match. For example, in selection of material for a new design, a computer can simultaneously search for a match of four, five or more parameters, each of which can be stated as below or above some target value or within a stated range. Just as quickly the effects on the extent of the candidate list or on their ranking of adding further requirements, relaxing others, or weighting the significance of each parameter in the required set are readily obtained. In another instance, materials properties data, fed directly into a design algorithm such as finite element analysis in the structural field, will give the final component design - geometry and dimensions - so that the effects of materials substitution can be immediately sensed and analyzed. In process automation, basic properties data from a computer file are combined with other property and process parameters, sensed dynamically in real time and fed to the microprocessor effecting the process automation. In other cases it is the ability of the computer to cope with huge volumes of data and to consider the simultaneous effects of several variables that is the feature exploited. Complex experimental programs or the analysis of data sets from multiple sources are cases in point. Many more examples of the unique advantages of computerized materials information could be adduced. In all cases, the computer system capabilities and applications described are not mere speculations but are present technical accomplishments. A recently prepared technical anthology (1) of published papers on computerized materials information systems well demonstrates this fact. The problem is that few of the several dozen materials properties databases built thus far (2) are publicly available, and all are small and narrowly focussed in subject coverage. Thus, while the applicability of the computer to materials information is well proven, an effective operating system of broad engineering utility is still lacking.

It is important to emphasize the materials information needs of engineers as opposed to scientists, for in many ways the needs of engineers in industry are more demanding and difficult to satisfy. Engineers must not only have accessible the best existing value for the property in question, but they must also know the limits of uncertainty for the value in order to estimate the reliability of a design. Rather than working solely with elemental metals or chemical compounds, engineers are frequently confronted with complex alloys, clads, and other composites whose properties and behavior must also be known or reliably estimated. Furthermore, many properties of commercial materials are not fixed values defined by pressure, temperature, and similar variables, but are history and structure dependent. In addition, engineers frequently require data that cannot be expressed in terms of a single property or combination of properties but must be related to a performance test or service application. There are fewer sources of compiled and evaluated data that cater to the needs of industry as compared to scientists generally; and under business pressures, engineers in industry have less time available to search for answers to their questions. Engineers, or their managers, are sometimes reluctant to release data to the outside world or to participate in cooperative projects because of a threat perceived (correctly or not) to their company's competitive position. Hence, much that is known and valuable does not become available to the world at large, much redundant work is undertaken, and real discrepancies are never examined and resolved. These considerations also underline the need and the opportunity for a revolutionary materials information system.

It may further be observed that the needs of the engineer user of a materials information system will vary accordingly as he is representative of a materials producer, a materials user, the R&D community or a developer of codes, standards and specifications. The specific information needs of these groups vary considerably. Materials producers are normally concerned with all properties of a narrow class of materials, while many materials users will confront only a limited set of properties or application regimes but wish to consider all available materials. Materials engineers from a user company, in either the design or manufacturing function, will require either nominal data for preliminary material or process selection or, on the other hand, specific data of stated reliability for design calculations. Note that the manufacturing engineering function normally includes selection of material for processability, control of manufacturing processes, and acquisition (purchase) of raw or semifinished material - each of which have their particular data requirements. Developers of codes, standards and specifications will wish to have full statistical representation and data analysis in order to prepare their documents and recommend routes to better quality assurance, life prediction and system performance. Table I repre-

	MATERIALS CLASSES									PROPERTIES GROUPS					
INDUSTRY GROUP	Ferrous	Non-Ferrous	Refractory & Superalloys	Ceramics & Glass	Composites	Inorganic Compounds	Plastics	Semi-Conductors	Wood	Mechanical	Thermal	Electrical, Electronic & Magnetic	Other Physical	Corrosion & Oxidation	Processability
civil engineering	•	•		•			•		•	•	•			•	
transportation	•	•	•		•		•			•	•			•	•
power generation	•	•	•							•	•	•		•	•
aerospace	•	•	•	•	•			•		•	•	•	•	•	•
defense	•	•	•	•	•	•				•	•	•	•	•	•
chemical	•	•		•						•	•			•	
oil and gas	•		•							•	•			•	
electronics and communication		•		•				•				•	•		•
materials producers	•	•	•	•			•		•	•	•	•	•	•	•
consumer products	•	•		•	•		•	•		•	•	•		•	•
other industrial products	•	•					•			•	•	•			•

Table I Materials Information Needs by Industrial Groupings.

sents an attempt to characterize the major data needs of various industrial groupings in terms of both materials and properties, while Table II, on the other hand, gives the different kinds of data required by materials producers and users in different functional areas.

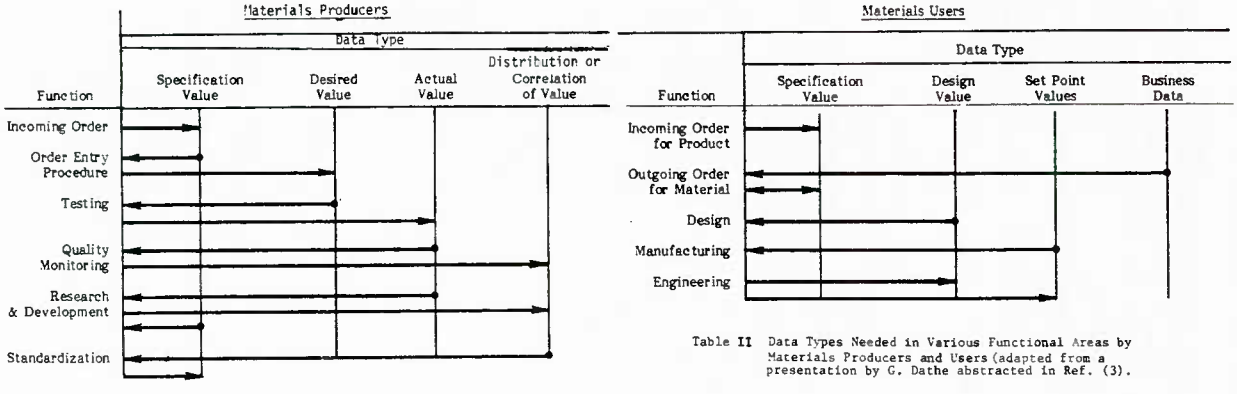


Table II Data Types Needed in Various Functional Areas by Materials Producers and Users (adapted from a presentation by G. Dathe abstracted in Ref. (3).

THE PRESENT STATE OF AFFAIRS IN THE MATERIALS DATA FIELD

Despite the obvious and multifaceted need for materials data in engineering and the growing intensity of that need, the situation remains unsatisfactory in many respects:

- Materials data sources are not widely known despite the recent publication of a series of critical surveys (4) and a general review of data sources for materials scientists and engineers (5).
- Gaps, contradictions, and inconsistencies in materials data have not been generally apparent because of the lack of a) thorough compilations and b) awareness of those that do exist.
- With some notable exceptions (e.g., the CINDAS program at Purdue University and MPC's efforts), critical evaluations of engineering properties of materials are generally missing.

- The lack of ready access to reliable materials data is causing a significant economic penalty as well as suboptimal performance of engineering systems.

- The power of the computer is not being exploited in the materials information field to the extent that it is in other business, scientific and technical fields.

- The leadership exerted by the U.S. government 15-20 years ago in materials information storage, analysis, and dissemination has now largely been lost.

Superficially, it would seem that modern computer technology has the potential to solve many of these problems as is proven by its success in many limited experiments and in the building of small private files alluded to above. However, in contrast to the successful development of computerized databases in the bibliographic, sociometric and financial fields, achievement and broad use of automated materials databases have been impeded by additional complexities and difficulties intrinsic to the materials engineering field. Among them are:

- *Data files must be created* ab initio

The data must be found and extracted from the public literature, private reports and other sources. They must then be evaluated and reliability estimates deduced. Similar data from different sources that are to be amalgamated must first be homogenized as to units, measurement conditions and other referrent facts. Finally, to be truly useful and to make the most efficient use of the computer, the data must be organized according to some schema of knowledge.

- *The development is people-limited*

Persons capable of building an effective materials database must be simultaneously knowledgeable of materials data and their applications and of modern computerized information technology. Ineffective and unused systems have now disappeared which were built by individuals or groups expert in but one of these two critical areas. Individuals or groups possessing both of the required backgrounds are rare indeed.

- *Property values must be associated with various descriptors*

Materials properties of engineering interest, aside from fundamental constants such as melting point or density, are invariably functions of other parameters which must also be included if the stored value is to have any real meaning. Among these are the common independent variables - temperature, pressure, section size etc, history and structure of the material, and the test method.

- *The information to be stored is varied*

There are facts (text), values (numbers) and relations (graphs or equations). To deal with all these types within a single information system poses special problems in data entry, search and retrieval.

- *Search operations are complex*

In addition to strict Boolean "and/or/not" matching, search routines must also cope with relational requirements (equal to, less than, greater than, within the range of) and permit extrapolation and interpolation from stored values.

- *Existing machine-readable materials data files have been little used and seem capable of generating only limited revenue*

This is a "Catch 22" situation. The limited use derives from the limited access and limited compatibility of the existing group of databases which in turn derive from the failure to make the investment of time and money to improve the content of the databases, to increase their scope coverage and to provide for their interconnection. Implicit to all this is the poor perception of the cost/benefits of ready access to reliable materials data.

SOME RECENT STUDIES

Two recent studies have examined the need for, feasibility of, timeliness and best mode of development of a coordinated system of databases covering the engineering properties of materials. The first, a year-long study by the writer, was commissioned by the Metal Properties Council, an umbrella organization of several professional societies with interest in materials data: American Society for Testing and Materials, American Society of Mechanical Engineers, American Society for Metals and American Welding Society. Their original premise was that the coordinated system envisaged might be brought into being through the cooperative efforts of several societies, principally those constituting the Metal Properties Council. The study that MPC commissioned was to:

- justify the need and define the user market
- identify immediate and future participants
- identify databases suitable for inclusion

- recommend a management and funding plan
- identify technical problems to be overcome
- define a materials-properties matrix suitable for a "Phase I" system
- define the search, display and manipulation capabilities appropriate to a "Phase I" system

The second study was conducted in a workshop mode by a group of materials data experts, assembled by a diverse steering committee, and was funded by support from the National Bureau of Standards, the Committee on Data for Science and Technology of the International Council of Scientific Unions, the Fachinformationszentrum of the Federal Republic of Germany and the Oak Ridge National Laboratory. In the workshop mode, small task groups addressed predefined questions in each of several topical areas: the feasibility, desirability, and timeliness of on-line access to materials data; attainable goals; formulation of a plan for implementation of a coordinated effort; and definition of unresolved problems for future attention.

In the event, both of these studies culminated at about the same time: the Workshop was held in November 1982, and the report to the Metal Properties Council was made in December 1982. (6) The formal report from the Workshop was published in the Spring of 1983. (3) The report from a third study conducted by the National Materials Advisory Board is yet to be issued.

CONCORDANT OBSERVATIONS FROM TWO STUDIES

A remarkable and highly reassuring result of the MPC study and the Materials Data Workshop was the high degree of consistency and agreement in their conclusions and recommendations. The more important of these concordant observations are as follows:

- *Computer access to engineering properties of materials is badly needed but does not now exist in any comprehensive way.*

Among the several drivers for computer accessible engineering data on materials are a) the advent of new technologies such as space and nuclear power, b) the thousands of new materials being brought to market and c) the new requirements imposed on materials for reliability, health and safety, and environmental protection. As engineers have turned either to their traditional printed sources of information [for summary listings see Refs (4) and (5)] or to the few developmental computerized databases [catalogued in Ref (6)], they have found them inadequate for the various reasons described above. Even taken as a group, the existing machine-readable databases do not serve, not only because of their non-existent inter-connection and compatibility, but also because of the gaps and redundancies in their coverage. Some sense of the latter may be inferred from the summary listing of Table III derived from the catalog of machine-readable databases previously published (6):

By Countries		By Material		By Property	
30	U.S.				
7	U.K.				
4	Japan				
3	France				
3	Canada				
2	W. Germany				
2	Poland				
1	E. Germany	39	Metals	40	Mechanical
1	Netherlands	4	Composites	7	Process-related
1	Italy	2	Semiconductors & Electronics	2	Thermal
1	Austria	3	Processing	2	Chemical
1	Jugoslavia	8	Polymers	2	Electrical & Electronic
1	Norway	1	Inorganic Compounds		
1	Czechoslovakia	1	Fibers	5	Other
58		58		58	

Table III Characteristics of Identified Machine-Readable Engineering Materials Properties Databases

- *Direct materials information needs and computerized engineering and manufacturing systems are driving this development.*

While the computer offers the possibilities of improved currency of information, better materials selection, and development of more nearly optimized materials, the current surging growth of computer-aided-design (CAD) and computer-aided-manufacturing (CAM) activities positively demand it. So far, CAD and CAM have largely addressed the problems of automating the drafting, calculational and geometric aspects of these functions. The reference data for insertion into design algorithms or into process controls must still be input by hand, largely from printed sources. As a consequence, the efficiencies of full automation are impaired, accuracy lost, and consistency undermined.

- *A cost-effective and technically sound computerized system of engineering properties data on materials will not arise spontaneously; a cooperative effort toward a distributed system of independent databases is needed.*

Attempts to build computerized materials properties databases can be traced back almost 20 years. None has been successful in the sense of a publicly available, comprehensive, widely used resource for the reasons previously discussed: limited scope, poor data reliability, inadequate promotion, difficulty of access and inconvenience in use. Current developments in computer technology - lower cost hardware and more powerful and versatile software - will not basically alter the situation. A cooperative effort by all interested stakeholders must be initiated. Such an endeavor, especially if in the form of a distributed, coordinated system of independent databases, as detailed elsewhere (7) would offer many advantages:

cost sharing
focussed expertise
decreased redundancy of effort
complementarity of needs, skills and resources
ability to marshall a critical level of effort
opportunity for standardization and compatibility
enhanced promotion and training

- *The participating stakeholder groups should include the professional societies, trade associations, universities, research institutes, industrial users and generators of materials data, technical publishers, and the government. They should be international in scope.*

Professional Societies. These societies should have a major role in setting standards on material identification, property definition, test standardization, and related matters. They should also establish the criteria for the inclusion of data and labeling of the level of validation. They can help to identify gaps in the data and educate users (their members) on new information technologies. It is likely that in order to provide convenient access to data for their members, the professional societies will produce and maintain one or more computerized databases within their specialties, just as they are now producing printed data compilations. Some could involve themselves directly in the electronic dissemination process. But a more likely scenario is for the societies to join in some form of consortium to create and operate the system as a whole.

Trade Associations. The role of trade associations is primarily that of a vehicle for raising funds for investment from their industry members and, in some cases, for obtaining data from member companies. Some may contribute databases to the system.

Private Industry. Industry is a source as well as a major user of the data that are to be incorporated in the databases. Industrial users are in the best position to determine priority needs for data and to set the required levels of reliability. Industrial materials producers are in the best position to contribute, via new process technology, to improved properties and more consistent properties, once a correlation between process and properties variation becomes established. It is also hoped that industry also will provide some of the seed money needed for establishing the system. Companies can further help in the initial stages by promoting use of the system by their engineers. It should also be noted that, while large industrial companies may be able to make the most immediate and most sophisticated use of computerized materials information systems, the greatest benefit may be to the host of smaller companies. The latter lack the knowledgeable staff to make proper materials decisions, the time to search the literature for data, and the financial and physical resources to generate necessary data themselves.

Information Vendors. This category includes conventional publishers as well as on-line vendors. Their role depends on the institutional arrangements that develop for setting up the system. Some publishers might contribute databases derived from their handbooks. Journal publishers may be able to channel unpublished data to societies or other centers that are preparing databases and may also be able to help by instituting data tagging and flagging programs. Information vendors may also be able to assist in dissemination.

Government. Government laboratories represent a large data generation and data user group. Also they can be approached as a major source of funding for establishing the databases and setting up the dissemination system. Any consortium building a comprehensive materials database will have to work with government to reduce artificial barriers such as antitrust restraints.

Universities. Universities will be a minor but still important source of data and data users. Academic research can make a significant contribution to a system of materials databases through a better understanding of materials properties. This research will permit more critical evaluation of the data. Educational institutions also have an important role in training engineers in the use of new data sources and information handling techniques.

International Organizations. CODATA, WFEO, ISO, UNIDO and other international organizations can provide mechanisms for setting standards and promoting international collaboration in building databases. Such international activity should not be restricted to international organizations per se but should also extend to multi-lateral cooperation between any interested individual nations.

- *No significant computer technology related barriers exist.*

Computer technology necessary for remote access to a distributed network of databases embraces three subtechnologies: hardware, software and telecommunications. Present state-of-the-art is quite adequate to support an acceptable materials information system. The principal, but surmountable, barrier to an effective distributed system at present is inadequate standardization, especially with regard to the means of designating materials and properties. Despite the present adequacy of computer technology for the subject purpose, significant changes in the near term can be confidently expected such as: new types of communication links, more powerful and "user-friendly" software, video storage for handling mixed text, graphic and numeric data, more cost-effective means of digitizing available printed copy, and options for distribution of custom-packaged data subset/software combinations for use on personal and other microcomputers. These, however, will supplement and enhance rather than totally replace the present computer/phone-line/CRT terminal in the immediate future.

- *A time window now exists for achieving an effective, coordinated information system.*

Several threats are impending which might prohibit realization of the desired system. Among these are:

- Emergence of one or more sub-professional quality databases prompted by commercial interests will, by a kind of Gresham's Law, inhibit or abort the growth of computerized materials data systems generally.

- Unilateral development, commercial or national, of an effective comprehensive system could establish a dangerous monopoly on quality materials information with consequent effects on the goods and services derivative from it.

- A Tower of Babel effect could come about lacking a coordinated development of mutually compatible files, languages, and database management systems. That is, over a short period of time so many independent databases could develop that bringing them together later in some kind of effective, rational functioning unit might become well-nigh impossible.

- *The immediate next step is to identify the best group to lead the development, raise financial resources and coordinate technical expertise.*

Required characteristics of the leadership group or "umbrella" organization include prestige, objectivity and political and societal connections such that access to funding and key personnel are eased. Candidate organizations in the U.S. include the National Academy of Engineering, the Numeric Data Advisory Board, the Metal Properties Council or a new ad hoc not-for-profit organization. Internationally, organizations such as CODATA or UNESCO might be considered. The required developmental financing is in the range of several million dollars - not a large sum as major technical projects go but beyond the capabilities of most organizations for unilateral action. Shared financing is thus another reason for a cooperative endeavor.

- *Realization of an effective, comprehensive and coordinated system of databases may be expected to have many synergistic and autocatalytic effects.*

  - Development of nationally (and internationally) accepted nomenclature, definitions, and properties.

  - Protocols for complete, detailed reporting of all data, the time of testing, measurement or evaluation to facilitate audit trails, reliability and validity assessment, and use. (This could become an essential, integral activity of standards groups, such as each of ASTM's technical committees, as additional practices supporting a standard.)

  - Protocols for consistent, uniform, and complete presentation of data in scientific and technical literature.

  - Increased attention to standardization of estimation protocols (mathematical models, graphics, etc.)

  - Additional and more uniform protocols for data appraisal or assessment of the reliability and validity of data.

  - Development of a computer-searchable reference database that would match specifically identified and completely described materials, conditions of testing, etc., and properties versus the literature source, laboratory that did the work, etc.

  - A master database, computer-searchable, that identifies test method versus source (on an international basis).

  - Evolution of a master matrix of data available versus data not available, for use in planning materials property data development programs.

- Evolution of a flexible national (and international) numbering (identification) system for all types of material for which data are being made available, controlled by a national (or international) organization (perhaps similar to the ISBN system for books).

- Evolution of user groups in each type of materials data use activity, in each industry, to help define specific data needs and help to support development and operation of databases.

- Establishment of additional, coordinating groups like MPC and DIPPR for cognizance over specific classes of properties.

- Move of the abstracting services towards more complete coverage of data, with data flagging and tagging.

- Increased participation by academia and others in making data available that reside in theses, dissertations, reports and books from the academic presses, laboratory files, etc.

- Concerted efforts by DOD and other government agencies to require their own laboratories and those of contractors to make data sets, data banks, or databases available for universal use.

- Increased participation by industry in contributing privately acquired data and in evaluating publicly available data.

SUBSEQUENT ACTIONS

During 1983, following the conclusion of the MPC study and the Materials Data Workshop, several actions were taken in support of their recommendations:

- The main findings of the MPC study as to the need, feasibility (both technical and economic), and distributed database concept for a comprehensive computerized materials information system were endorsed in principle by the oversight committee and subsequently approved by MPC's Board of Trustees.

- The Metal Properties Council announced (8) the formation of a new not-for-profit corporation, the National Cooperative Materials Information Network to design, fund and manage a new system. Formation of a board of trustees for the new corporation is now in progress.

- A separate study (9) by a National Materials Advisory Board Committee endorsed the findings of both the MPC study and the Materials Data Workshop report. This study particularly emphasized the key role that U.S. government agencies can and should take in bringing the envisioned information system to realization.

- The discussions, conclusions and recommendations of the Materials Data Workshop were summarized in a substantial report (3).* In accordance with one of these recommendations, copies of the report were forwarded to the president of the National Academy of Engineering with the request that the findings be reviewed by pertinent bodies within NAE and appropriate actions taken.

- The background and findings of both studies reviewed here have been publicized by lectures and papers presented to such groups as ASME, ASIS, CODATA, AIChE (10) and the Design Engineering Conference.

- A workshop, "Towards a National Science and Technology Data Policy", was held April 14, 1983 which highlighted the U.S. need for a national data policy in science and technology, noting the present disarray and inadequacy of existing U.S. data programs. The workshop was coordinated by the Committee on Science and Technology of the U.S. House of Representatives, the Congressional Research Service of the Library of Congress and the Numerical Data Advisory Board of the National Academy of Science.

- Plans have been made to convene, during 1984, three smaller application oriented workshops, each involving primarily representatives of a single industrial area, viz:

  ground vehicles
  power generation
  aerospace and defense

  The purpose of these workshops would be to more carefully define and prioritize data needs and secure commitments of funding and technical assistance in building the planned system.

*Copies of this report are available to qualified individuals upon application to Dr. J.R. Rumble, Jr., Office of Standard Reference Data, National Bureau of Standards, Washington, DC 20234

- Several small individual tasks have been initiated to maintain momentum and build interest in the concept of a comprehensive, coordinated system of quasi-independent databases, pending the formation of a permanent leadership organization. Among these activities are:
  - Representatives of the MPC and the NBS have been meeting with the leadership and key technical committees of various professional societies attempting to define particular roles and tasks for each in the total endeavor.
  - A detailed description of the scope, content, status, access mode, and computer specifications of all known machine-readable databases on the engineering properties of materials is being compiled by means of a questionnaire survey.
  - A study is being initiated of the standardization required with respect to materials, properties and test method descriptors so as to permit effective interaction between separate databases in a distributed system. This work will concentrate on the mechanical properties of metals.

SUMMARY

After several years of independent, uncoordinated activity in the computerization of data on the engineering properties of materials, it is now realized that cooperation by a diverse group of stakeholders will be required to achieve an effective, comprehensive system. We are just now at the point of identifying the appropriate organization to lead such an effort and of distributing the various tasks which have to be done to those groups with the interest, resources and skills to carry them out. Achievement of the goal of an effective operating system by the end of the decade seems attainable.

REFERENCES

1. Westbrook, J.H. and Rumble, J.R., "Selected Readings on Computerized Materials Data Systems". An anthology for the Materials Data Workshop, Tennessee, Nov 1982, 468 pp.

2. Hampel, V.E., Hilsenrath, J., Westbrook, J.H., Gaynor, C.A., and Johnson, P.S. "A Directory of Databases for Material Properties", Lawrence Livermore Laboratory Report UCAR 10099, Aug (1983).

3. Westbrook, J.H. and Rumble, J.R., "Computerized Materials Data Systems," Proceedings of the Materials Data Workshop, Fairfield Glade, TN (1982) 133 pp.

4. a) Gavert, R.B., Moore, R.L., Westbrook, J.H. 1974, "Critical Surveys of Data Sources: Mechanical Properties of Metals. NBS Special Publication 396-1,81pp. SD Catalog No. C13.10:396-1.

   b) Johnson, D.M., Lynch, J.F. 1975, "Critical Surveys of Data Sources: Properties of Ceramics." NBS Special Publication 396-2,47 pp. SD Catalog No. C13.10:396-2.

   c) Diegle, R.B., Boyd, W.K. 1976. "Critical Surveys of Data Sources: Corrosion of Metals." NBS Special Publication 396-3,29 pp. SD Catalog No. C13.10:396-3

   d) Carr, M.J., Gavert, R.B., Moore, R.L., Wawrousek, H.W., Westbrook, J.H. 1976. "Critical Surveys of Data Sources: Electrical and Magnetic Properties of Metals." NBS Special Publication 396-4.83 pp. SD Catalog No. C13:10:396-4.

5. Westbrook, J.H. and Desai, J.D., "Data Sources for Materials Scientists and Engineers," Ann. Review of Materials Science 8 (1978) 359 pp.

6. Westbrook, J.H., "Feasibility Study of An Inter-Society Computer-Based Materials Property System", Report to the Metal Properties Council, Dec 1982, 91 pp.

7. Westbrook, J.H., "Cooperation in Developing Computerized Materials Data Bases." Data for Science and Technology. Proceedings of 8th International CODATA Conference, Oct. 1982, Jachranka, Poland. North-Holland Publ. Co. (1983) 91-98 pp.

8. MPC News Release, "MPC Announces Plan for National Cooperative Materials Property Data System Based on Coordinating Existing Facilities", May 4, 1983.

9. Brown, W.F., et al, "Materials Information Used in Computerized Design and Manufacturing Processes" NMAB Report-405 (1982).

10. Graham, J.A., "The Metal Properties Council's Activities on a National Materials Property Data Network" AIChE meeting, Denver, Colorado, August 1983.

ACRONYMS

AIChE	American Institute of Chemical Engineers
ASIS	American Society for Information Science
ASME	American Society of Mechanical Engineers
ASTM	American Society for Testing and Materials
CINDAS	Center for Information and Numerical Data Analysis and Synthesis (Purdue University)
CODATA	Committee on Data for Science and Technology (ICSU)
CRT	Cathode Ray Tube
DIPPR	Design Institute for Physical Property Data (of AIChE)
DOD	Department of Defense (U.S.)
ICSU	International Council of Scientific Unions
ISBN	International Standard Book Number
ISO	International Standards Organization
MPC	Metal Properties Council (U.S.)
NBS	National Bureau of Standards (U.S.)
UNESCO	United Nations Educational, Scientific and Cultural Organization
UNIDO	United Nations Industrial Development Organization
WFEO	World Federation of Engineering Organizations

DATA ORGANISATIONS AND THEIR MANAGEMENT

Dr John R Sutton

Scientific and Technical Information Unit
Department of Trade and Industry
Ebury Bridge House
Ebury Bridge Road London SW1W 8QD

ABSTRACT

Organisations which are involved in generating, compiling, validating and disseminating data are not all alike. Different types of organisations have different objectives and motivations. These lead to differences in management. Ways of coordinating the activities of data organisations are considered and the scope for overall planning at national and international levels. The costs of data activities cannot be ignored. The economics of subsidies, pump-priming and pricing need careful consideration.

1 INTRODUCTION

Data organisations differ and they differ in a number of ways. They may be independent or part of a larger organisation, profit or non-profit seeking, public or private sector. They may be dependent, or partly dependent on external funding - from government or otherwise. They may differ in how many and which of the stages of the data cycle they are involved in and in the availability of their results to users outside the organisation.

2 PRIVATE SECTOR DATA ORGANISATIONS

A data organisation which is financially as well as legally independent of larger bodies needs to sell its output at a price which covers all of its expenditure including staff, accommodation, marketing and the acquisition of whatever constitutes the starting point for its operations. It is unlikely that such an organisation would originate data by measurement - except perhaps by providing measurement services for clients. Compilation is a natural activity for such an organisation as it provides an opportunity to add value without incurring high costs. Validation is often associated primarily with experts at large public organisations (universities or government laboratories) but it is sometimes argued that validation should be carried out independently of data generation and a small organisation might employ experts who had previously developed their expertise and reputation in such large organisations. To be viable an independent data organisation will need to find a niche supplying some type of data which its users value sufficiently and dissemination, including marketing, will be an important activity. Data organisations which are part of a private sector enterprise may take the form of subsidiary companies employing their own staff but are more likely to operate as an internal service team. The users will certainly include, and may be predominantly, other staff of the parent company. In the larger companies and groups there may be either several data teams associated with the operating divisions of the company or

an integrated central team associated with a research capability - or perhaps both. The data organisation's output may be distributed freely inside the company or available on request and in the larger companies there may be some form of notional charging against budgets which although expressed in money terms are somewhat less 'real' than external expenditure. There are several possible graduations in the availability of data outside the parent company. The output might be available only to closely related organisations such as companies within the group, companies in which group companies have a substantial financial interest or with whom they trade extensively. It might be available to anyone, or almost anyone who asks either free or for a fairly nominal charge - perhaps calculated to cover the extra cost incurred in providing extra copies of documents, computer tapes or discs etc. Or there might be a positive effort to disseminate the output and to recover a significant fraction of the costs while still treating in-house users as the primary users whose existence justifies the investment in the data team and whose needs determine the programme of work. This last case could even lead to the formation of a more or less independent subsidiary dealing at approximately 'arms length' with the rest of the group, and expected to make a profit.

Data organisations within companies may include the generation of data in their range of activities but this is more likely in the case of large companies or groups which have research laboratories. Small and medium sized enterprises are less likely to devote resources to data generation. Compilation of data from external sources is likely to be the predominant actvity in this type of data organisation. In the larger companies the data team may include staff who have become recognised experts in a particular area of data and validation will be an important activity. The extent and nature of the team's work on data dissemination will depend on the extent to which data is made available outside the organisation, as referred to above, and also on the customary style of intracompany communications. In some cases a comprehensive company databook or databank will be compiled and updated at regular intervals. In other cases compilations may be produced for each major project - for example an aircraft engine or a plant to manufacture a particular chemical.

3 ACADEMIC DATA ORGANISATIONS

Academic data organisations will usually have developed out of a research team involved in generating data and compiling data from other sources initially for comparison with their own data. If the team is also developing theories to explain the variations of a property from one substance to another or with temperature, composition, time etc then the compilation will also be used to test the theories. If there is a lack of already compiled data for the particular area of interest then copies of the teams compilations may be requested by other researchers in the field leading to publication of the compilation and growing involvement in compilation and validation. The team's

research expertise may be in a highly specific aspect of data generation and the available manpower largely research students spending only a few years in the team. Transition to a fully fledged data operation is likely to involve increased staff and perhaps also seeking funds through different channels. The peer review committees involved in grant applications for academic research may consider data compilation less worthy of support because it appears to be less intellectual, less original, less academic than the generation of original data, the development of a new method of measurement or a new theory. A data centre properly needs funding for a longer period than the conventional grants linked to the training period of a PhD research student.

4 GOVERNMENT LABORATORIES

Data organisations in government laboratories differ according to whether the laboratory or the agency it is part of, is mission-oriented or not. If it is then there is a vast appetite for data relevant to that mission and the data organisation's task is to satisfy as much as possible of that appetite. Committees and other means of communication with the users of each type of data can be used to establish programmes and priorities. In-house use predominates together with use by other organisations involved in the agency's activities whether as suppliers, subject to regulation etc. The data output may be made available to others, subject to security considerations in the case of defence matters. Once it is established that the data is needed by some part of the agency then the cost of providing it is a necessary part of the cost of the agency - provided the data operations are conducted efficiently and without waste.

The activities of this kind of data organisation will include the generation of data whenever this is more efficient than buying in data from outside. The data required might be highly specialised or needed at short notice or it may be desirable to keep close control over the data generation. Compilation will be included and validation in the areas where agency staff are the experts. In other areas external experts might be employed in validation as consultants or contractors. Dissemination inside the organisation and maintaining good contacts with the user groups will be important activities but external dissemination is secondary and may not justify assignment of scarce resources.

Data organisations in government laboratories conducting fundamental research or research intended to support the country's industrial base may provide an equally fundamental support to the availability of general data capable of a wide range of use. Generation of data is likely to play a large part in the activities of such organisations but the types of data generated will be determined by the research programme of the laboratory. When an area of research reaches the stage where the fundamentals are well established and the detailed application can best be carried out closer to the user - either in industry or in mission oriented agencies, then the laboratory's programme should be phased down to allow resources to be concentrated in newer areas. While the laboratory is active in generating data of a particular type it may be one of the major locations of expertise in that subject on a national or even an international

scale. As such it is a natural focus for the compilation and validation of data. The data will be made widely available and it is desirable that dissemination should be a carefully considered activity with adequate resources.

5 OTHER DATA ORGANISATIONS

Data organisations may also be operated as part of the activities of a professional institution. The members will not wish to see their subscriptions used to support such an activity unless the support is modest and the benefits to all or most of the members sufficiently evident. Funds may be found from elsewhere and the operation may avoid running at a loss. One of the strengths of a professional institution is its ready access to knowledgeable people in that profession including both academics and senior management in industry.

6 ENGINEERING SCIENCES DATA UNIT

Several UK data organisations will now be considered. The Engineering Sciences Data Unit grew out of work carried out at the Royal Aeronautical Society to meet a need for aircraft design data during World War II. This work was wholly funded by government. The work continued after the war with government and industry funding and in 1965 expanded to include design data of interest to mechanical and chemical engineers. Government funding was provided as launching aid through the Institution of Mechanical Engineers and the Institution of Chemical Engineers. At the end of the launching period the Unit became self supporting from the sale of its output. At this time the Unit took the form of a company, wholly owned by the Royal Aeronautical Society, and dissemination and marketing was carried out by a subsidiary of the Thomson Organisation - a private sector company with considerable experience in the information sector. More recently the Unit has also become a subsidiary within the Thomson Organisation. The chairman, Dr A J Barrett, is a lecturer in this lecture series and has been involved in ESDU throughout the development from being part of a professional institution to its present position.

The Unit compiles data from the published literature and additional material obtained through industrial, academic or government laboratory staff serving on a number of technical committees. Validation by the staff of the unit in interaction with the technical committee is an important part of the Unit's work. The chief output is a series of some 800 printed Data Items covering mechanical, structural, aeronautical and chemical engineering and also materials data. An increasing number of computer programs are also produced but the data is considered to be unsuitable for online access.

Dissemination of the Data Items is by sale either on a standing order basis or as single items. Marketing involves visits to users and potential users to promote sales and identify needs. User feedback arises chiefly from the regular meetings of the technical committees. About half of the items are sold to users outside the UK.

7 PHYSICAL PROPERTY DATA SERVICE

This data service was set up at the Institution of Chemical Engineers in the mid 1970's. The service was supported financially by the Department of Industry on a launching aid basis. The first manager of the Service was an existing member of the Institution staff who had been closely involved in the proposal to set up the service. The second was seconded from a large chemical company for two years. This proved a useful way for him to broaden his experience and for the Service to benefit from someone with direct knowledge of the factors affecting the use of the data in industry. A management committee, chaired by a senior industrialist and with members from industry, university and government department was supported by technical committees or panels on which users and data experts were represented. Recently the Service has been restructured to combine the marketing skills developed at the Institution of Chemical Engineers with the data generating and validating expertise of the National Engineering Laboratory of the Department of Trade and Industry.

The data available to the Service consisted initially of a collection of data on the physical properties of a considerable number of chemicals. This had been compiled in a large chemical company, initially for internal use, and later sold to the Service. Additional data was obtained from government laboratories and other sources and computer software to manipulate the databank, including the calculation of the properties of mixtures, was also obtained - largely from another chemical company. The government laboratories also played an important role in validating the data. Links with organisations involved in the development of computer software for the design of chemical process plant proved important in the dissemination of the data for use with such design software.
The data output consisted initially of a single databank available online. Then a combination of databank and access software was made available for purchase or lease. Later a number of separate packages of interest to different users were made available.

8 MASS SPECTROMETRY DATA CENTRE

This centre was set up in 1966 at the Atomic Weapons Research Establishment (AWRE) of the UK Atomic Energy Authority. A survey of the requirements of mass spectrometrists had shown a need for such a centre to provide a collection of spectra and thereby assist the identification of unknown substances. The Centre was funded initially by the Office of Scientific and Technical Information of the Department of Education and Science and subsequently by the Ministry of Technology and the Department of Industry. An increasing proportion of the costs of the centre was covered by income from sales of the Centre's output. Changes at AWRE, including its transfer to the Ministry of Defence led to the Centre moving to become part of the Information Services Section of the Royal Society of Chemistry (RSC). This part of the RSC is located on the campus of

Nottingham University in leased accommodation. The Department of Industry provided financial support for a period after this transfer but the Centre is now independent of external support.

The Centre identifies input material by searching relevant journals and abstracts. Authors of papers are asked to donate their spectra to the collection and arrangements are also made to incorporate collections of data. The input data is inspected and checked before entering the collection. A major bibliographic output is the monthly Mass Spectrometry Bulletin with comprehensive indexes. This bibliographic information is available on magnetic tape and also online through Pergamon Infoline. Spectra were originally available on data sheets but these were superseded by magnetic tape. The data is also available online through the NIH/EPA Chemical Information System. The spectra are also available in condensed form in the Eight Peak Index. This is a multi volume reference book now in its third edition.

In the earlier years the printed output was printed and distributed by a government agency and limited resources were available for marketing. This is now carried out by the RSC marketing team who were already experienced in marketing information services including Chemical Abstracts. Users and potential users are widely distributed in industry, universities and government and in many different areas of science and technology including pharmaceuticals, medicine, pollution and effluent monitoring, quality control and forensic science. Contact with representative users is provided by regular meetings of a technical advisory committee.

9 CRYSTALLOGRAPHIC DATA CENTRE

This Centre was set up in 1965 at the University of Cambridge and has remained there ever since. The Director, Dr Kennard, has been involved in the centre from its inceptio It is funded partly by government and partly by income from dissemination of the output. At first the government funding was from the Office of Scientific and Technical Information in the Department of Education and Science. In 1974 this office became part of the British Library but its data activities passed to the Science Research Council (SRC) now renamed Science and Engineering Research Council (SERC). A new Data Committee was set up by the SRC to handle these data activities for a transitional period after which responsibility for the Centre passed to the Chemistry Committee. The Centre compiles data from published papers and from unpublished additional supplementary data but does not itself produce any original measurements. Additional information is added in the form of derived data - eg chemical connectivity - also checking and recalculation are used to validate the data. The chief output is in the form of computer tapes, issued annually with four-monthly updates, of three files - a bibliographic file, a chemical connectivity file and a file of structure data.

Dissemination is through about twenty affiliated centres each serving a country or group of countries. The database is leased to these centres for use within their country or countries. The centres are supported by grants from their national governments and make a contribution to the Centre proportionate to the number of crystallographers in their territory. In the UK the data is available through the regional computer centres and the online networks of the SERC without charge except for communication costs. Online access is also available in the US and France but not through commercial or other networks which cross national boundaries. This restriction is necessary because of the territorial pattern of dissemination. The principal users are crystallographers, pharmaceutical chemists and chemists concerned with catalysis. Communication with users who are also authors of papers occurs in the process of checking and validation. Conferences and workshops provide additional contact with users.

10 LESSONS FROM THESE DATA CENTRES

It is clear from the brief accounts of these centres that real life is in many ways less 'tidy' than the idealised general cases. Much depends on individuals who know what they want to achieve and are determined to find a way through the constraints required by the policy of government or other funding bodies.

Government agencies may be restructured, divided or merged and their areas of responsibility redefined or redistributed. A data centre receiving financial support from such an agency may find that different criteria now govern the administration of that support. Some of the differences will be fairly trivial - the use of different forms to report progress or claim reimbursement. Others will be more significant - for example a different view of the importance of market forces, of the appropriate timescale for financial support or of the appropriate balance between national and international considerations.

Changes in the economic climate can lead to reductions in financial support, in the availability of new measurements and in the demand, and ability to pay for the output. Reductions in financial support may be accommodated by increased income from sales which may require diversion of resources to strengthen the sales team, by trimming costs either by increasing efficiency or by eliminating or postponing the less essential activities, by finding additional support from alternative sources or by some combination of these. Temporary reductions in the availability of new measurements do not diminish a centre's claim to include all, or almost all, of the best data in its field - indeed such reductions ease the task of maintaining the data collection in an up to date state. If the organisation's users have reduced demands or are less able to afford data the impact on the organisation will depend on the dissemination arrangements. If the data is sold on a price per item basis the impact may be immediately significant. If sales are on a subscription or bulk sale basis then some users may be reluctant to renew their subscriptions and new users will be harder to find. Apart from increased effort from the sales team the organisation may need to accept some loss of income and concentrate its efforts on being ready to respond when the business cycle reaches its turning point. There may be scope to expand into new areas.

A data organisation will not necessarily remain at the same location, either in the geographical or the organisational sense. The examples above include a transition from a professional body to the private sector, from a government laboratory to a professional body and from a professional body to a government laboratory (albeit with a continuing role for the professional body). Such changes may be a consequence of changes in funding or other economic factors or they may arise from changes in the nature and organisation of the original parent body.

It is possible for a data centre to develop from an initial stage heavily dependant on grants to a position of greater independence. This depends on good marketing. It is necessary, but not sufficient, to apply resources - both expenditure and capable manpower - to promote sales. It is also necessary to learn who the users are, how they use the data, how they could use it and to incorporate this knowledge into the decisions on what to produce and how to do it.

Generating new data by measurement is usually the most expensive part of the data cycle unless it is a byproduct of some other activity. Compilation and validation are less expensive and can be more easily funded by selling the output. Adding value by deriving additional data from the input, by constructing a comprehensive databank or by providing the means of manipulating the data to match a users particular problem can make a very significant contribution to the success of a data organisation.

Data organisations need to be ready to adapt to changes such as the shift from printed output to computer oriented forms of output. There are differences here between data and bibliographic or textual information. The latter are intended to be read by the user and a display at a computer terminal, although in some ways less convenient, can be made acceptable to the user. Numerical data are likely to be needed by the user for some form of calculation. This calculation, unless it is a very simple one, will be performed on a computer and the data organisation is supplying one component to a calculation process controlled by the computer software. There will usually be other components of input to the process which the user will obtain from other sources or provide himself and the process may well require interaction with the user so that he can take decisions that depend on preliminary results from the calculations. The design of the software is therefore likely to be dominated by factors relating to the nature of the calculations and even more the nature of the interaction with the user. The data organisation may be able to influence the software design in order to facilitate the transfer of data but will often have to accept and accommodate to constraints and formats laid down by the software designer. In some cases it may be possible for the data organisation to move into the design of software. Users in different sectors of the data organisation's market may differ in the speed at which they wish or are willing to move to computer oriented data usage.

11 TECHNICAL COMMITTEES

Most data organisations make use of committees on which users are represented as well as those involved in generation, compilation and validation. The number of users who are members of any such committee is necessarily limited and it is important that they can

take a wider view - representing the interests of other users in their sector, discipline or project - as well as expressing personal views on and needs for data. One way to include more users is to have a series of committees for the various identifiable disciplines, subdisciplines etc. These committees need effective chairmen to steer the committees away from being talking shops and towards constructive contributions to the assignment of priorities and the hard choices between activities competing for limited resources. In the case of data organisations producing data primarily for inhouse use or other non-commercial distribution these committees should determine the programme of work by a process of interaction between users, who know what they need, and data experts, who know what is possible. In the case of data organisations operating more commercially it will be necessary for final decisions to be taken by those responsible for the commercial operation of the organisation but the technical committees should still assign technical priorities and be allowed to play a significant part in the process. Mutual respect between the two groups will be a key factor in success.

The technical committees should also be looking ahead to spot trends in data usage and try to assess the future needs. This is difficult and will never produce predictions which turn out to be absolutely correct but it is important that the organisation should collect together the best estimates it can from its users. Senior managers in the user organisations may be able to help in taking a somewhat longer view than those who are fully occupied with immediate problems.

12 NATIONAL AND INTERNATIONAL COORDINATION

There are advantages in some coordination of data activities on larger scales - especially the national and international scales. Such coordination is, of course, not an attempt to determine the detailed policies and plans of all the independent bodies involved in data activities. Some degree of such coordination is provided by the major funding bodies in their consideration of a variety of applications. In the UK these have included the Requirements Boards of the Department of Trade and Industry and the subject committees of the SERC. Although SERC had, for a few years, a separate Data Committee applications for support for academic data activities are now considered by the appropriate subject committee for chemistry, physics etc. This has the advantage of bringing data compilation activities together with data measurement, development of new measurement methods and others aspects of research in that subject.

The British National Committee on Data for Science and Technology is a Royal Society committee as the Royal Society is the UK adhering body to CODATA. Many of the countries represented on CODATA also have national committees and apart from advising the national representative on matters pertaining to CODATA they usually have a role as a national coordinating committee for data. Representatives from the major national bodies such as, in the UK, Deparment of Trade and Industry, SERC, British Library together with academics and industrialists involved in data can consider general issues and encourage cooperation and awareness of data activities.

At the international level the Committee on Data for Science and Technology (CODATA) is a committee of the International Council of Scientific Unions and has both national members and members from the Scientific Unions (for Chemistry, Physics etc). CODATA supports a number of international working groups on various aspects of data, has developed recommendations on the presentation of data, guides to data activities throughout the world, training programmes on data activities etc.

13 COSTS, PRICES AND SUBSIDIES

The costs of a data organisation are not always brought together in full. There are two main reasons for this. A data organisation which is part of a larger organisation may not be required to pay, or to pay a full economic price, for certain services provided by the parent organisation. There are a number of such services and policy on which to charge for will differ in different organisations. When notional charges are calculated for management purposes they may have the opposite effect of burdening the data organisation with a share of the overheads of the parent organisation which may need different facilities.

The other factor affecting costs is the provision of inputs as a contribution in kind. The market for such input data is often a limited one and a fair price difficult to determine. Contribution in kind, perhaps associated with privileged or earlier access to output or discounted prices for the output, avoids the problem.

Nevertheless costs are incurred when data is generated - especially if this involves measurements of high precision requiring specialised equipment in expensive laboratories. Usually these costs will be carried as part of the cost of a research laboratory. In industry this will be justified by the need to develop new products and the defensive need to be aware of and to counter factors - such as pollution or product liability - which may diminish the company's profitability. Government research costs will usually be justified as an infrastructure contribution to national industrial capabilities and in some cases by strategic considerations.

Compilation costs are smaller. They can be, and for data used outside the parent organisation usually are, covered by charges for the output data. The cost of validation depends partly on how deep the process of checking or reassessment goes. It can be a modest component within a total for compilation and validation or reach a substantial fraction of the original measurement cost. If modest it can be recovered with the compilation cost. Extensive validation is usually associated with data required inhouse and justified by its value to the parent organisation. There may be scope for partial cost recovery from external sales.

The costs of dissemination depend on the difficulty of reaching the target users but will usually be related to the target income from sales - within a limit determined by market saturation.

Prices will often be limited by the user's present willingness to pay. In the past much data was available very cheaply and users may need encouragement to appreciate the value of well compiled and validated data. The user's internal costs in obtaining the data for himself are not always fully visible for comparison with the price of data from a data centre. Much will depend on the skill of the data centre in choosing the type of data to produce and in packaging and presenting it. Accepting input data in exchange for output or as part payment may help.

A new data centre may need external support for an initial period because, until the products are developed and its reputation established with the users, income from sales is unlikely to cover costs. This pump-priming or launching aid type of support allows a new centre to come into existence when it would not otherwise do so. A possible alternative would be to subsidise, for a temporary period, the user's purchase of the centre's products. This would normally be more difficult to administer and less effective. In either case the funding body will need to consider carefully the probable viability of the centre after the launching period.

Long term support for a data centre is harder to justify in economic terms. The expected benefits need to be carefully assessed against the maximum benefit achievable by alternative use of that expenditure. In the case of government expenditure that would include the benefits to the national economy of not taking that sum out of the productive sector. If the output of a data centre is largely used by users who are supported by the same funding body and expenditure in support of the data centre is replacing part of the support which would otherwise have been needed by the users then support of the data centre may be cost effective.

14 CONCLUSIONS

There is no simple uniform pattern for the setting up and management of data organisations. It is necessary to adopt a flexible approach and to treat each individual case on its merits. People are important and data projects like many others often depend for their success on some individual with characteristics not altogether unlike a successful entrepeneur. Administrative and procedural requirements - especially in large organisations - are not always as bad as they are often represented. The barriers needed to keep order can become hurdles for an athletic expression of fitness.

Data organisations need good technical ideas and also sound marketing. Resources, people and financial, must be available for both of these - and the people must be able to work together. So long as the organisation has both technical and marketing strength either could play the leading role and either might be the determining factor in the organisation's location in the wider community - whether industrial, academic, government or otherwise.

Market forces are a major feature of the developed, non-centralised economies of countries in NATO, EEC, OECD etc. When the unrestrained operation of short term market forces is perceived to have longer term disadvantages then temporary subsidies, launching aid or pump-priming may be necessary. In some cases support on a longer timescale may be justified.

ACKNOWLEDGEMENTS

This text draws on the experience of my colleagues in the Department of Trade and Industry to whom I express my thanks.

BIBLIOGRAPHY

This Bibliography with Abstracts has been prepared to support AGARD Lecture Series No.130 by the Scientific and Technical Information Branch of the National Aeronautics and Space Administration, Washington, D.C., in consultation with the Lecture Series director Dr R.F.Taschek.

UTTL: A bibliography - Remote sensing as related to change detection
AUTH: A/MATHEWS, M.; B/MILLER, L. D. PAA: B/(Texas A&M University, College Station, Tex.) Remote Sensing Quarterly, vol. 2, Apr. 1980, p. 47-52.
ABS: A list of 41 technical documents pertaining to the use of remote sensing techniques in the analysis and mapping of changes in urban and natural vegetation land use is presented. The list was compiled by the retrieval of the relevant documents from the RESENA (remote sensing of nature) computerized bibliography of approximately 16,000 references on remote sensing, by the use of a computer key-word-in-title retrieval system. 80/04/00 81A10159

UTTL: CODATA Directory of Data Sources for Science and Technology. I - Crystallography' CODATA Bulletin, no. 24, June 1977, 48 p.
ABS: This chapter is primarily concerned with structural, property and physical data derived from or used in crystallographic studies. The format adopted includes the following sections: Section A - international data projects; Section B - national data projects; Section C data centers; Section D - major publication series; Section E - other data sources; and Section F - bibliography. Indexes are included by subject and by country in order to facilitate information retrieval. 77/06/00 80A34020

UTTL: Bibliographic and numeric data bases for fiber composites and matrix materials'
AUTH: A/MCMURPHY, F. E.; B/QUICK, T. M. PAA. B/(California, University, Livermore, Calif.) In: ICCM/2; Proceedings of the Second International Conference on Composite Materials, Toronto, Canada, April 16-20, 1978. (A79-16981 05-24) Warrendale, Pa., Metallurgical Society of AIME, 1978, p. 33-43.
ABS: The Data Management Group at the Lawrence Livermore Laboratory is conducting research leading to the creation of data bases for energy storage systems. These data bases are computer-based and will contain bibliographic information, material properties data, and data on essential criteria for energy storage systems. Access to these central files will be from remote terminals over computer networks and by telephone dialup, in addition to the more conventional means of computer-generated reporting, and dissemination on magnetic tapes. Bibliographic and numerical data bases have been created for fiber composites and matrix materials, with particular emphasis on their application to modern flywheel technology. 78/00/00 79A16984

UTTL: Database management systems for numeric/structured data, an overview
AUTH: A/HAMPEL, V. E. CORP: California Univ., Livermore. Lawrence Livermore Lab. CSS: (Integrated Information Systems.) Presented at the 7th Intern. CODATA Conf., Kyoto, 8-11 Oct. 1980
RPT#: DE82-001681 UCRL-84594 CONF-801095-3 80/10/00 83N71929

UTTL: Linking bibliographic data bases: A discussion of the Battelle technical report
AUTH: A/JONES, C. L. CORP: Council on Library Resources, Inc., Washington, D.C.
RPT#: ED-195-274 IR-009-053 80/10/15 82N77941

UTTL: Synfuels data base for stationary-combustor applications
AUTH: A/FRIGO, A. A.; B/CLINCH, J. M.; C/FISCHER, J. CORP: Argonne National Lab., Ill. CSS: (Energy and Environmental Systems Div.)
RPT#: DE82-010398 ANL/EES-TM-164 81/09/00 82N77059

UTTL: Processes for application of protective coatings: Diffusion coatings. Citations from the Metals Abstracts data base CORP: National Technical Information Service, Springfield, Va.
RPT#: PB82-860313 82/01/00 82N77007

UTTL: Medical information systems. Citations from the INSPEC data base CORP: National Technical Information Service, Springfield, Va.
RPT#: PB81-866824 81/05/00 82N76897

UTTL: Lapping techniques for the preparation of polished surfaces. Citations from the Metals Abstracts data base CORP: National Technical Information Service, Springfield, Va.
RPT#: PB81-872046 81/07/00 82N76725

UTTL: Sheet metal joining: Ferrous alloys. Citations from the Metals Abstracts data base CORP: National Technical Information Service, Springfield, Va.
RPT#: PB82-858903 81/12/00 82N76724

UTTL: Protective coatings for pipelines. Citations from the metals abstracts data base CORP: National Technical Information Service, Springfield, Va.
RPT#: PB82-858564 81/12/00 82N76581

UTTL: The tensile properties and machinability of leaded brasses. Citations from the metals abstracts data base CORP: National Technical Information Service, Springfield, Va.
RPT#: PB81-872202 81/07/00 82N76579

UTTL: The wear characteristics and lubrication of journal bearings. Citations from the metals abstracts data base CORP: National Technical Information Service, Springfield, Va.
RPT#: PB81-872194 81/07/00 82N76578

UTTL: Flotation separation of metallic sulfide particles. Citations from the metals abstracts data base CORP: National Technical Information Service, Springfield, Va.
RPT#: PB82-858739 81/12/00 82N76577

UTTL: Flotation separation of nickel particles. Citations from the metals abstracts data base CORP: National Technical Information Service, Springfield, Va.
RPT#: PB82-858846 81/12/00 82N76576

UTTL: Electrodeposition of copper base alloys. Citations from the metals abstracts data base CORP: National Technical Information Service, Springfield, Va.
RPT#: PB82-859075 81/12/00 82N76574

UTTL: Flotation separation of copper particles. Citations from the metals abstracts data base CORP: National Technical Information Service, Springfield, Va.
RPT#: PB82-859364 81/12/00 82N76573

UTTL: Scientific and technical, spatial, and bibliographic data bases the US geological survey, 1979 CORP: Geological Survey, Reston, Va.
RPT#: USGS-CIRC-817 LC-80-600027 CIRC-817 79/00/00 82N70991

UTTL: Data base languages. Citations from the NTIS data base CORP: National Technical Information Service, Springfield, Va.
RPT#: PB81-807711 NTIS/PS-79/1010 81/06/00 82N70866

UTTL: CODATA Directory of Data Sources for Science and Technology. Chapter 5: Seismology, no. 42
AUTH: A/EMPTOZ, G.; B/LIDE, D. R., JR.; C/WESTRUM, E. F., JR.; D/LANDER, J. F. CORP: International Council of Scientific Unions, Paris (France). CSS: (Committee on Data for Science and Technology.) Pergamon Press, Ltd.
RPT#: ISSN-0366-757X 81/06/00 81N77365

UTTL: A directory of computer software applicatons. Library and information sciences CORP: National Technical Information Service, Springfield, Va.
RPT#: PB80-228463 NTIS/SA-80/05 80/10/00 81N76138

UTTL: Relational data bases. Citations from the NTIS data base
AUTH: A/JONES, J. E. CORP: National Technical Information Service, Springfield, Va.
RPT#: PB81-800427 NTIS/PS-78/1086 80/11/00 81N73757

UTTL: Data base languages. Citations from the NTIS data base
AUTH: A/CARRIGAN, B. CORP: National Technical Information Service, Springfield, Va.
RPT#: PB80-809866 NTIS/PS-79/1010 NTIS/PS-78/0977 80/04/00 81N71417

UTTL: Copyrights. Citations from the NTIS data base
AUTH: A/YOUNG, M. E. CORP: National Technical Information Service, Springfield, Va.
RPT#: PB80-810823 NTIS/PS-79/0574 NTIS/PS-78/0529 80/06/00 81N70746

UTTL: Data base languages. Citations from the NTIS data base
AUTH: A/CARRIGAN, B. CORP: National Technical Information Service, Springfield, Va.
RPT#: NTIS/PS-79/1010/2GA NTIS/PS-78/0977 NTIS/PS-77/0878 79/10/00 80N75280

UTTL: Relational data bases. A bibliography with abstracts
AUTH: A/CARRIGAN, B. CORP: National Technical Information Service, Springfield, Va.
RPT#: PB80-800725 NTIS/PS-78/1086 NTIS/PS-77/0896 79/11/00 80N73424

UTTL: Spatial data on energy, environmental, socioeconomic, health and demographic themes at Lawrence Berkeley Laboratory, 1978 inventory
AUTH: A/BURKHART, B. R.; B/MERRILL, D. W. CORP: California Univ., Berkeley. Lawrence Berkeley Lab.
RPT#: LBL-8744 79/04/00 80N72695

UTTL: Computer information security and protection, volume 2. Citations from the NTIS data base
AUTH: A/CARRIGAN, B. CORP: National Technical Information Service, Springfield, Va.
RPT#: NTIS/PS-79/0865/0 NTIS/PS-78/0860 NTIS/PS-77/0629 NTIS/PS-76/0562 NTIS/PS-75/437 79/09/00 80N71496

UTTL: Bibliographic and numeric data bases for fiber composites and matrix materials
AUTH: A/MCMURPHY, F. E.; B/QUICK, T. M. CORP: California Univ., Livermore. Lawrence Livermore Lab. Presented at Flywheel Technol. Symp., San Francisco, 5 Oct. 1977
RPT#: UCRL-79503 CONF-771053-2 77/09/00 80N70441

UTTL: Data base management. Citations from the Engineering Index data base
AUTH: A/CARRIGAN, B. CORP: National Technical Information Service, Springfield, Va.
RPT#: NTIS/PS-79/0385/9 NTIS/PS-78/0329 NTIS/PS-77/0315 NTIS/PS-76/0266 79/05/00 79N78230

UTTL: Data base management. Citations from the NTIS data base
AUTH: A/CARRIGAN, B. CORP: National Technical Information Service, Springfield, Va.
RPT#: NTIS/PS-79/0384/2 NTIS/PS-78/0328 NTIS/PS-77/0314 NTIS/PS-76/0265 79/05/00 79N78229

UTTL: Sanitary landfills. Citations from the Engineering Index data base
AUTH: A/SMITH, M. F. CORP: National Technical Information Service, Springfield, Va.
RPT#: NTIS/PS-78/1185/4 NTIS/PS-77/1049 NTIS/PS-76/0820 78/11/00 79N74011

UTTL: Relational data bases. A bibliography with abstracts
AUTH: A/REIMHERR, G. W. CORP: National Technical Information Service, Springfield, Va.
RPT#: NTIS/PS-78/1086/4 NTIS/PS-77/0896 78/10/00 79N71435

UTTL: Development of a gridded data base. Appendix A: The 3DNEPH data base. Appendix B: Analysis data base summary. Appendix C: The usefulness of the gridded conventional data base for climatic application
AUTH: A/FEDDES, R. G. CORP: Air Force Environmental Technical Applications Center, Scott AFB, Ill.
RPT#: AD-A056234 USAFETAC-TN-74-2 74/04/00 79N70557

UTTL: ERDA bibliographic data base
AUTH: A/CAPE, J. D. CORP: Energy Research and Development Administration, Oak Ridge, Tenn. CSS: (Technical Information Center.)
RPT#: CONF-750336-1 75/00/00 78N77681

UTTL: Data base languages. Citations from the engineering index data base
AUTH: A/GROOMS, D. W. CORP: National Technical Information Service, Springfield, Va.
RPT#: NTIS/PS-78/0012/1 78/01/00 78N74674

UTTL: Computerized scheme for duplicate checking of bibliographic data bases
AUTH: A/GILES, C. A.; B/BROOKS, A. A.; C/DOSZKOCS, T.; D/HUMMEL, D. J. CORP: Oak Ridge National Lab., Tenn. CSS: (Computer Sciences Div.)
RPT#: ORNL/CSD-5 76/08/00 78N74183

UTTL: Wind shear. Citations from the engineering index data base
AUTH: A/HABERCOM, G. E., JR. CORP: National Technical Information Service, Springfield, Va.
RPT#: NTIS/PS-77/1173/2 NTIS/PS-76/1041 77/12/00 78N74025

UTTL: Data base languages. A bibliography with abstracts
AUTH: A/GROOMS, D. W. CORP: National Technical Information Service, Springfield, Va.
RPT#: NTIS/PS-77/0878/7 77/10/00 78N71986

UTTL: Relational data bases. A bibliography with abstracts
AUTH: A/GROOMS, D. W. CORP: National Technical Information Service, Springfield, Va.
RPT#: NTIS/PS-77/0896/9 77/10/00 78N71676

UTTL: Data bases and data base systems related to NASA's Aerospace Program: A bibliography with indexes
CORP: National Aeronautics and Space Administration, Washington, D. C.
ABS: This bibliography lists 641 reports, articles, and other documents introduced into the NASA scientific and technical information system during the period January 1, 1981 through June 30, 1982. The directory was compiled to assist in the location of numerical and factual data bases and data base handling and management systems.
RPT#: NASA-SP-7048 NAS 1.21:7048 83/00/00 83N22010

UTTL: Research into the structure, accessing and manipulation of numeric data bases. Report on phase 1
AUTH: A/LISTON, D. M., JR.; B/WIEDERKEHR, R. V.; C/KING, D. W.; D/SCHRAM, P.; E/GOUDY, K.; F/DOLBY, J. L.
CORP: King Research, Inc., Rockville, Md.; Dolby Associates, Los Altos, Calif.
ABS: The report covers Phase 1 of a 3-Phase research and development program to produce a commercially marketable system for the efficient storage, retrieval, manipulation, and publishing of numeric data. The Phase 1 research explores the technical feasibility of a system concept which employs counterpart systems: (1) a data system which gathers, stores, retrieves, manipulates, and publishes numeric data, and (2) its counterpart metadata system which stores, retrieves and manipulates metadata (information about the numeric data). The research also explores the use of faceted classification as a mechanism for detailed indexing of numeric data based on a body of theory being developed by James L. Dolby growing out of research on energy data handling.
RPT#: PB82-238056 KRI-9338 82/03/00 83N19647

UTTL: Longwall data bank CORP: Bituminous Coal Research, Inc., Monroeville, Pa.
ABS: Progress is reported in compiling and transferring to the coal industry comprehensive longwall operational data representing 95% of the US longwall installations as well as presenting abstracts of domestic and foreign literature published since 1975 related to longwall production, productivity, and dust control. The literature search was completed, except for an ongoing review of weekly abstract publications. Working cards were completed for the comprehensive index file for all longwall publications of interest; and abstracts were prepared for selected longwall publications. Revised questionnaire forms, with information from the original questionnaires, were forwarded for updating to nearly all companies presently in the questionnaire file. Selected operational data were tabulated and compared.
RPT#: DE82-005090 DOE/FE-00080/2 BCR-L-1260 81/12/00 82N29679

UTTL: Non-bibliographic online database: An investigation into their uses within the fields of economics and business studies
AUTH: A/HOUGHTON, B.; B/WISDOM, J. C. CORP: British Library Lending Div., Boston Spa (England).
ABS: A state of the art survey on the use of nonbibliographic data bases in Great Britain was conducted. Special attention was given to the fields of economics and business studies.
RPT#: PB82-149618 ISBN-0-905984-70-0 ISSN-0308-2385 REPT-5620 81/00/00 82N28211

UTTL: Technology transfer and the management of human services CORP: National Clearinghouse for Improving the Management of Human Services, Rockville, Md.
ABS: The Project SHARE bibliography on technology transfer in the human services sector is an information resource selected to reflect recent research, demonstrations, transfer experiments, and experience in the human services field. It is a sample of the documentation on technology transfer, not an exhaustive listing of the literature in the field or in the Project SHARE data base. As such, it should be an aid to determining what technologies are available for transfer in human services as well as how transfer might be effectively implemented.
RPT#: PB82-105644 81/08/00 82N18077

UTTL: Non-bibliographic online database: An investigation into their uses within the fields of economics and business studies
AUTH: A/HOUGHTON, B.; B/WISDOM, J. C. CORP: British Library Lending Div., Boston Spa (England).
ABS: A state-of-the-art survey of nonbibliographic online database services is presented. Emphasis is on the potential of access to nonbibliographic online databases and the development of teaching packages to introduce potential users to nonbibliographic online databases.

RPT#: BLL-BLRDR-5620 ISBN-0-905984-70-6 ISSN-0308-2385 81/05/00 81N34084

UTTL: Data bases and data base systems related to NASA's aerospace program. A bibliography with indexes
CORP: National Aeronautics and Space Administration, Washington, D. C.
ABS: This bibliography lists 1778 reports, articles, and other documents introduced into the NASA scientific and technical information system, 1975 through 1980.
RPT#: NASA-SP-7045 81/06/00 81N31018

UTTL: Data bases available at the National Bureau of Standards library
AUTH: A/CUNNINGHAM, D. CORP: National Bureau of Standards, Washington, D.C.
ABS: An alphabetical listing either by acronym or full title of the data base is presented. Additional information provides description of the data base, period of coverage, producer(s), corresponding hard copy, principal sources and vendors. A general subject and a cross reference index to the data bases is also supplied.
RPT#: PB81-132870 NBSIR-80-2133 80/10/00 81N20955

UTTL: A broad knowledge of information technologies: A prerequisite for the effective management of the integrated information system
AUTH: A/LANDAU, H. CORP: Midwest Research Inst., Golden, Colo. Presented at the 12th Ann. Meeting of the Western Canada Chapter of ASIS, Saskatchewan, 25 Sep. 1980
ABS: There is a trend towards the bringing together of various information technologies into integrated information systems. The managers of these total systems therefore must be familiar with each of the component technologies and how they may be combined into a total information system. To accomplish this, the manager should first define the overall system as an integrated flow of information with each step identified; then, the alternate technologies applicable to each step may be selected. Methods of becoming technologically aware are suggested and examples of integrated systems are discussed.
RPT#: SERI/TP-750-926 CONF-800986-1 80/09/00. 81N16949

UTTL: Distributed data processing. Citations from the Engineering Index data base
AUTH: A/CARRIGAN, B. CORP: National Technical Information Service, Springfield, Va.
ABS: The bibliography of worldwide journal literature cites studies on the concepts, design, development, implementation, and application of distribution data processing. Also included are studies on distributed data bases. General communication studies related to major computer networks are cited in another bibliography. This updated bibliography contains 243 abstracts, none of which are new entries to the previous edition.
RPT#: PB80-812233 NTIS/PS-79/0711 NTIS/PS-78/0672 80/06/00 80N32282

UTTL: Distributed Data processing. Citations from the Engineering Index data base
AUTH: A/CARRIGAN, B. CORP: National Technical Information Service, Springfield, Va.
ABS: The bibliography of worldwide journal literature cites studies on the concepts, design, development, implementation, and application of distributed data processing. Also included are studies on distributed data bases. General communication studies related to major computer networks are cited in another bibliography. This updated bibliography contains 231 abstracts, all of which are new entries to the previous edition.
RPT#: PB80-812225 NTIS/PS-79/0711 NTIS/PS-78/0672 80/06/00 80N32281

UTTL: Distributed data processing. Citations from the NTIS data base
AUTH: A/CARRIGAN, B. CORP: National Technical Information Service, Springfield, Va.
ABS: The bibliography of Federally-funded research cites studies on the concepts, design, development, implementation, and application of distributed data processing. Also included are studies on distributed data bases. General communication studies related to major computer networks are cited in another bibliography. This update bibliography contains 164 abstracts, 37 of which are new entries to the previous edition.
RPT#: PB80-812217 NTIS/PS-79/0710 NTIS/PS-78/0671 80/06/00 80N32280

UTTL: Computer science and technology: Data Abstraction, databases, and conceptual modelling. An annotated bibliography
AUTH: A/BRODIE, M. L. CORP: Maryland Univ., College Park. CSS: (Dept. of Computer Science.)
ABS: A bibliography containing 350 entries on issues within the area of conceptual modelling of dynamic systems of complex data is given. The entries are from recent work in the areas of database management, programming languages, artificial intelligence, and software engineering.
RPT#: PB80-183833 NBS-SP-500-59 LC-80-600052 80/05/00 80N30055

UTTL: Chemicals identified in human biological media: A data base, volume 1, part 2, records 1-1580
AUTH: A/CONE, M. V.; B/BALDAUF, M. F.; C/MARTIN, F. M.; D/ENSMINGER, J. T. CORP: Oak Ridge National Lab., Tenn. CSS: (Health and Environmental Studies Program.)
ABS: A comprehensive data base of chemicals identified in human biological media (tissues and body fluids) is presented in two volumes. Introductory material, references, appendices, indices, and a chemical directory are given in this volume as a user guide to the data base.
RPT#: ORNL/EIS-163/V1-P2 APR-1 80/03/00 80N29030

UTTL: Chemical identified in human biological media: A data base, volume 1, part 1, records 1-1580
AUTH: A/CONE, M. V.; B/BALDAUF, M. F.; C/MARTIN, F. M.; D/ENSMINGER, J. T. CORP: Oak Ridge National Lab., Tenn. CSS: (Health and Environmental Studies Program.)
ABS: A comprehensive data base of chemicals identified in human biological media (tissues and body fluids) is presented in two volumes. The data base in given in this volume in tabular form and arranged alphabetically by CAS 'preferred chemical name'.
RPT#: ORNL/EIS-163/V1-P1 APR-1 80/03/00 80N29029

UTTL: Data base management. Citations from the NTIS data base
AUTH: A/CARRIGAN, B. CORP: National Technical Information Service, Springfield, Va.
ABS: The advent of on-line systems, and the increasing problems of file organization, file maintenance, and file structures of data bases, has required the study and development of data base management systems. This bibliography of Federally funded research cites the development of software packages and implementation of data base management systems into various information systems. Also cited are guidelines for use in optimizing and modelling data bases. This updated bibliography contains 255 abstracts, none of which are new entries to the previous edition.
RPT#: PB80-808157 80/04/00 80N28230

UTTL: Data base management. Citations from the Engineering Index data base
AUTH: A/CARRIGAN, B. CORP: National Technical Information Service, Springfield, Va.
ABS: The advent of on-line systems and the increasing problems of file organization, file maintenance, and file structures of data bases have resulted in the study and development of data base management systems. This bibliography of worldwide literature cites research on the development of software packages and the implementation of data base management systems into various information systems. This updated bibliography contains 251 abstracts, 66 of which are new entries to the previous edition.
RPT#: PB80-808173 NTIS/PS-79/0385 NTIS/PS-78/0329 80/04/00 80N28229

UTTL: Data base management. Citations from the NTIS data base
AUTH: A/CARRIGAN, B. CORP: National Technical Information Service, Springfield, Va.
ABS: The advent of on-line systems and the increasing problems of file organization, file maintenance, and file structures of data bases has required the study and development of data base management systems. This bibliography of Federally funded research cites the development of software packages and implementation of data base management systems into various information systems. This updated bibliography contains 96 abstracts, 87 of which are new entries to the previous edition.
RPT#: PB80-808165 NTIS/PS-79/0384 NTIS/PS-78/0328 80/04/00 80N28228

UTTL: Alcohol fuels, volume 2. Citations from the NTIS data base
AUTH: A/CAVAGNARO, D. M. CORP: National Technical Information Service, Springfield, Va.
ABS: Federally-funded research on alcohol based fuels that may be used in the future as a fuel source is presented. Synthesis, chemical analysis, performance testing, processing, pollution, economics, environmental effects, and feasibility are included. One hundred and thirty five abstracts, 109 of which

are new entries to the previous edition are reported.
RPT#: NTIS/PS-79/0713/2 NTIS/PS-78/0673 NTIS/PS-77/0620 79/07/00 79N33342

UTTL: Energy information data base. Serial titles, supplement 3 CORP: Department of Energy, Oak Ridge, Tenn. CSS: (Technical Information Center.)
ABS: Changes and additions to TID-4579-R10 (the authority list for serial titles used by TIC) are contained in this supplement. The supplement is intended to be used with that publication.
RPT#: TID-4579-R10-SUPPL-3 78/12/00 79N30088

UTTL: SBIBLI: A data base system for bibliographic references
AUTH: A/SCHUBERT, L. CORP: Forschungsinstitut fuer Funk und Mathematik, Werthoven (West Germany).
ABS: A program system is presented to aid the prospective user in creating, updating, and above all, using his individual file of bibliographic references. The system is thought to bridge the gap between the personal card catalog and the huge, powerful, and not always attainable documentation system. It is highly user-oriented. The user is not assumed to be intimately familiar with bibliographies or information retrieval. The functions of the system are described in some detail and its structure and design sketched. How to use the system is thoroughly explained and examples are given.
RPT#: FFM-264 78/03/00 79N28052

UTTL: Data base languages. Citations from the Engineering Index data base
AUTH: A/CARRIGAN, B. CORP: National Technical Information Service, Springfield, Va.
ABS: Query languages, data definition languages, data manipulation languages, and data dictionary/directories were studied. References are made to the CODASYL standard, as well as to many other data bases. The data elements of data bases applied to a variety of areas are described. This updated bibliography contains 60 abstracts 20 of which are new entries to the previous edition.
RPT#. NTIS/PS-79/0004/6 NTIS/PS-78/0012 79/02/00 79N21818

UTTL: A study to review fire prevention and control innovations and new technologies developed by Federal agencies since January 1, 1971. Part 2: Compendium key word listing
AUTH: A/JOHNSON, M. A.; B/FOTHERGILL, J. W.; C/BRYAN, J. L.; D/DOUBERLY, E. B. CORP: Integrated Systems, Inc., Rockville, Md.
ABS: The Key word is listed with cross references to document numbers that appear as citations with abstracts in Part 1: it is essentially a key work index to about 2900 documents resulting from federally sponsored projects.
RPT#: PB-287384/2 REPT-09A78000378 77/11/30 79N16158

UTTL: A study to review fire prevention and control innovations and new technologies developed by Federal agencies since January 1, 1971. Part 1: Compendium
AUTH: A/JOHNSON, M. A.; B/BRYAN, J. L.; C/FOTHERGILL, J. W.; D/DOUBERLY, E. B. CORP: Integrated Systems, Inc., Rockville, Md.
ABS: Literature citations are presented with abstracts of approximately 2900 documents that were cataloged and abstracted. The primary goal of the project was to create a resource data base on federally-sponsored work and to identify high value candidate technological and/or innovative work that could be immediately employed to improve fire safety. Approximately 3800 solicitations for information on such work were mailed over the course of the project in addition to an extensive literature search.
RPT#: PB-287383/4 REPT-09A78000377 77/11/30 79N16157

UTTL: A study to review fire prevention and control innovations and new technologies developed by Federal agencies since January 1, 1971. Volume 2: Thesaurus
AUTH: A/JOHNSON, M. A.; B/BRYAN, J. L.; C/FOTHERGILL, J. W.; D/DOUBERLY, E. B. CORP: Integrated Systems, Inc., Rockville, Md.
ABS: The thesaurus was prepared as a supportive tool for the computer-compatible information base developed for the study of fire prevention and control innovations and new technologies. It was used in the abstracting and citation operation. The thesaurus in combination with the computer-compatible information base and an adequate computerized information system will allow a user to construct keyword search profiles to retrieve all citations of interest in a given field of fire prevention and control.
RPT#: PB-287382/6 REPT-09A78000376 77/11/30 79N16156

UTTL: Experiment in data tagging in information-accessing services containing energy-related data CORP: Chemical Abstracts Service, Columbus, Ohio.
ABS: The use of data tags in a machine-readable output file was considered for incorporation into an on-line search service. Data tags are codes which uniquely identify specific types of numerical data in the corresponding source documents referenced in the file. Editorial and processing procedures were established for the identification of data types; the recording, editing, verification, and correction of the data tags; and their compilation into a special version of ENERGY, a CAS computer-readable abstract text file. Possible data tagging plans are described and criteria for extended studies in data tagging and accessing are outlined.
RPT#: PB-286654/9 CAS-264 78/08/00 79N15831

UTTL: Random access memories. Citations from the engineering index data base
AUTH: A/CARRIGAN, B. CORP: National Technical Information Service, Springfield, Va.
ABS: A bibliography containing 286 abstracts concerning research on the design, development, and application of random access computer storage devices is presented. Research on charge coupled devices in random access storage, random access memories in microprocessor applications, and reliability of random access memories are included.
RPT#: NTIS/PS-78/1222/5 NTIS/PS-77/1111 NTIS/PS-76/0998 78/11/00 79N15672

UTTL: Random access memories. Citations from the NTIS data base
AUTH: A/CARRIGAN, B. CORP: National Technical Information Service, Springfield, Va.
ABS: A bibliography containing 27 abstracts concerning research reports on the design, development, and application of random access storage devices and materials is presented. Studies are included on wall placement techniques, domain wall observations, microcircuitry, and software development.
RPT#: NTIS/PS-78/1221/7 NTIS/PS-77/1100 NTIS/PS-76/0997 NTIS/PS-75/811 78/11/00 79N15671

UTTL: Roller bearings. Citations from the engineering index data base
AUTH: A/HABERCOM, G. E., JR. CORP: National Technical Information Service, Springfield, Va.
ABS: A bibliography containing 214 abstracts concerning the design and performance of roller bearings is presented. Rolling contact bearing surfaces, hardening methods, mechanical properties, and lubricating techniques are among the areas reviewed.
RPT#: NTIS/PS-78/1111/0 NTIS/PS-77/0987 NTIS/PS-76/0887 78/10/00 79N14395

UTTL: Data base language. A bibliography with abstracts
AUTH: A/REIMHERR, G. W. CORP: National Technical Information Service, Springfield, Va.
ABS: The bibliography cites studies on query languages, data definition languages, data manipulation languages, and data element directories/dictionaries. References are made to the CODASYL standard, as well as to many other data bases. Descriptions of the data elements of data bases, applied to a variety of areas, are given.
RPT#: NTIS/PS-78/0977/5 NTIS/PS-77/0878 78/09/00 79N12792

UTTL: Alternatives for accessing engineering numerical data, phase 1
AUTH: A/VEYETTE, J. H., JR.; B/BEZILLA, R.; C/TOWLOWKIAN, Y. S. CORP: Engineering Index, Inc., N.Y.; Benson and Benson, Inc., Princeton, N. J.; Purdue Univ., Lafayette, Ind.
ABS: The data gathering procedures described ranged from utilization of primitive ad hoc methods to a mention of use of an in-house evaluated data system. For the most part, however, the engineer's information sources and procedures appeared to be confined largely to use of handbooks, departmental data, and personal contact with co-workers, in-house specialists, sub-contractors and vendors. The engineers surveyed seemed receptive to recognizing shortcomings in their current pragmatic practices. It was concluded that a proposal to make evaluated data available to the engineering community in on-line interactive mode will meet a set of true needs.
RPT#: PB-282609/7 EIX-78/1 78/03/00 79N10940

UTTL: Computer information security and protection. Citations from the engineering index data base
AUTH: A/REIMHERR, G. W. CORP: National Technical Information Service, Springfield, Va.
ABS: The bibliography of worldwide research includes the various aspects of computer information security and computer privacy, including personal privacy, reliability of security procedures, natural disasters, audits, electronic crime, and software design for efficiency checks. (This updated bibliography contains 182 abstracts, 24 of which are new entries to the previous edition.)
RPT#: NTIS/PS-78/0861/1 NTIS/PS-77/0630 NTIS/PS-76/0563 78/08/00 78N33765

UTTL: Computer information security and protection, volume 2. Citations from the NTIS data base
AUTH: A/REIMHERR, G. W. CORP: National Technical Information Service, Springfield, Va.
ABS: The bibliography of Federally-funded research covers the various aspects of computer information security and computer privacy, including personal privacy, reliability of security procedures, natural disasters, audits, electronic crime, implications of the Privacy Act of 1974, and software design for efficiency checks. (This updated bibliography contains 50 abstracts, all of which are new entries to the previous edition).
RPT#: NTIS/PS-78/0860/3 NTIS/PS-77/0629 NTIS/PS-76/0562 NTIS/PS-75/437 78/08/00 78N33764

UTTL: Computer information security and protection, volume 1. Citations from the NTIS data base
AUTH: A/GROOMS, D. W. CORP: National Technical Information Service, Springfield, Va.
ABS: The bibliography of Federally-funded research covers the various aspects of computer information security and computer privacy, including personal privacy, reliability of security procedures, natural disasters, audits, electronic crime, implications of the Privacy Act of 1974, and software design for efficiency checks. The military intelligence aspects of computer privacy and the private social implications are also included. (This updated bibliography contains 350 abstracts, none of which are new entries to the previous edition).
RPT#: NTIS/PS-78/0859/5 78/08/00 78N33763

UTTL: Cryogenic properties of aluminum and aluminum alloys. Citations from the Engineering Index data base
AUTH: A/SMITH, M. F. CORP: National Technical Information Service, Springfield, Va.
ABS: This updated bibliography contains 92 abstracts, 17 of which are new entries to the previous edition. Worldwide research on aluminum and its alloys in liquefied gas tanks, superconducting devices, pressure vessels, and spacecraft components are cited. Studies on welds, fracture, and mechanical properties are included.
RPT#: NTIS/PS-78/0512/0 NTIS/PS-77/0504 NTIS/PS-76/0354 78/06/00 78N32252

UTTL: Cryogenic properties of aluminum and aluminum alloys. Citations from the NTIS data base
AUTH: A/SMITH, M. F. CORP: National Technical Information Service, Springfield, Va.
ABS: This updated bibliography contains 134 abstracts, 21 of which are new entries to the previous edition. Citations of Federally-funded research include studies on the cryogenic properties of aluminum and its alloys used in superconducting machinery, magnets, space technology, and nuclear reactors. Electrical properties, fatigue, deformation, and welds are included.
RPT#: NTIS/PS-78/0511/2 NTIS/PS-77/0503 NTIS/PS-76/0353 78/06/00 78N32251

UTTL: Current knowledge on numerical data indexing and possible future developments CORP: Informatics, Inc., Rockville, Md.
ABS: A comprehensive review and analysis are given of programs aimed at developing methodologies for indexing the numerical data that occur in the literature used by scientists and engineers. Based on the study results, eleven recommendations are made which focus attention on the principal issues requiring consideration in any future data indexing efforts including the need for improved guidelines for indexes, more cooperative programs, and, ultimately, increased standardization.
RPT#: PB-279924/5 78/04/00 78N31953

UTTL: Microcomputers. Part 4: Process control applications. Citations from the engineering index data base
AUTH: A/HABERCOM, G. E., JR. CORP: National Technical Information Service, Springfield, Va.
ABS: Industrial applications of microcomputers to expedite

processes were investigated in these reports gathered in a worldwide literature survey. Such diversified industries as pulp mills, rubber processing industry, chemical plants, metal fabrication plants, and machine shops were researched to ascertain the effectiveness of microcomputers as a process control tool. Comparative analyses of microcomputers and their control systems are made.
RPT#: NTIS/PS-78/0616/9 78/06/00 78N31779

UTTL: Microcomputers. Volume 4, part 3: Basic design and development. Citations from the engineering index data base
AUTH: A/REIMHERR, G. W. CORP: National Technical Information Service, Springfield, Va.
ABS: The bibliography of worldwide research literature cites studies on the design and development of microcomputers. Studies on software development, chip development, LSI technology, and reliability and performance evaluation testing are included. (This updated bibliography contains 263 abstracts, all of which are new entries to the previous edition.)
RPT#: NTIS/PS-78/0615/1 NTIS/PS-77/0329 NTIS/PS-78/0610 78/06/00 78N31778

UTTL: Microcomputers. Volume 3, part 3: Basic design and development. Citations from the engineering index data base
AUTH: A/REIMHERR, G. W. CORP: National Technical Information Service, Springfield, Va.
ABS: The bibliography of worldwide research literature cites studies on the design and development of microcomputers. Studies on software development, chip development, LSI technology, and reliability and performance evaluation testing are included. (This updated bibliography contains 236 abstracts, none of which are new entries to the previous edition.)
RPT#: NTIS/PS-78/0614/4 78/06/00 78N31777

UTTL: Microcomputers. Volume 3, part 2: Telecommunication applications. Citations from the engineering index data base
AUTH: A/REED, W. E. CORP: National Technical Information Service, Springfield, Va.
ABS: The bibliography of worldwide research literature cites studies in telecommunication applications of microcomputers. Telephone, data transmission, teleprinters, facsimile communication, and communications controllers are among the applications cited. (This updated bibliography contains 91 abstracts, 48 of which are new entries to the previous edition.)
RPT#: NTIS/PS-78/0613/6 NTIS/PS-77/0328 NTIS/PS-76/0204 78/06/00 78N31776

UTTL: Microcomputers. Volume 4, Part 1: General applications. Citations from the engineering index data base
AUTH: A/REIMHERR, G. W. CORP: National Technical Information Service, Springfield, Va.
ABS: The bibliography of worldwide research literature cites studies on the applications of microcomputers. These include automotive, testing, navigation, instrumentation, biomedical, machine tools, and other applications. (This updated bibliography contains 370 abstracts, all of which are new entries to the previous edition.)
RPT#: NTIS/PS-78/0612/8 NTIS/PS-77/0327 NTIS/PS-76/0204 78/06/00 78N31775

UTTL: Microcomputers. Part 1: General applications. Citations from the engineering index data base
AUTH: A/REIMHERR, G. W. CORP: National Technical Information Service, Springfield, Va
ABS: The bibliography of worldwide research literature cites studies on the applications of microcomputers. These include process control, testing, navigation, instrumentation, biomedical, machine tools, and other applications. (This updated bibliography contains 191 abstracts, none of which are new entries to the previous edition.)
RPT#: NTIS/PS-78/0611/0 78/06/00 78N31774

UTTL: Microcomputers: General applications. Citations from the NTIS data base
AUTH: A/REIMHERR, G. W. CORP: National Technical Information Service, Springfield, Va.
ABS: This bibliography of Federally-funded research cites studies on the applications of minicomputers. These include process control, testing, navigation, instrumentation, biomedicine, machining, training, as well as other applications.
RPT#: NTIS/PS-78/0609/4 78/06/00 78N31773

UTTL: Infrared techniques for nondestructive testing. Citations from the engineering index data base
AUTH: A/HABERCOM, G. E., JR. CORP: National Technical Information Service, Springfield, Va.
ABS: A bibliography containing 90 abstracts concerning the fundamental principals of nondestructive testing and inspection, by use of infrared devices, is presented

Tire flaws, electronic circuit defects, and flaws in bonded surfaces are among the applications researched.
RPT#: NTIS/PS-78/0359/6 78/04/00 78N29485

UTTL: Infrared techniques for nondestructive testing. Citations from the NTIS data base
AUTH: A/HABERCOM, G. E., JR. CORP: National Technical Information Service, Springfield, Va.
ABS: A bibliography containing 106 abstracts concerning the fundamental principles of nondestructive testing and inspection, by use of infrared devices, is presented. Tire flaws, electronic circuit defects, and flaws in bonded surfaces are among the applications researched.
RPT#: NTIS/PS-78/0358/8 78/04/00 78N29484

UTTL: Gunn effect devices, volume 2. Citations from the engineering index data base
AUTH: A/REED, W. E. CORP: National Technical Information Service, Springfield, Va.
ABS: An updated bibliography containing 223 abstracts concerning the application of Gunn effect devices to microwave generation, amplification, and control is presented. The Gunn in semiconductors is discussed. Design, fabrication, and properties of Gunn diodes are included.
RPT#: NTIS/PS-78/0362/0 NTIS/PS-77/0257 78/04/00 78N29389

UTTL: Gunn effect and transferred electron devices, volume 2. Citations from the NTIS data base
AUTH: A/REED, W. E. CORP: National Technical Information Service, Springfield, Va.
ABS: An updated bibliography containing 64 abstracts concerning the application of Gunn effect and transferred electron devices to microwave generation, amplification, and control is presented. The Gunn effect in semiconductors is discussed. The design, fabrication, and properties of Gunn diodes and transferred electron devices are included.
RPT#: NTIS/PS-78/0361/2 NTIS/PS-77/0255 NTIS/PS-76/0271 NTIS/PS-75/227 78/04/00 78N29388

UTTL: Phased arrays. Citations from the engineering index data base
AUTH: A/REED, W. E. CORP: National Technical Information Service, Springfield, Va.
ABS: An updated bibliography containing 243 abstracts concerning reports from worldwide research on the design, performance, radiation patterns, and applications of phased arrays is presented. Applications include radar, communications, optical, electronic countermeasures, acoustic, aircraft, and spacecraft
RPT#: NTIS/PS-78/0332/3 NTIS/PS-77/0310 NTIS/PS-76/0309 78/04/00 78N29387

UTTL: Phased arrays, volume 3. Citations from the NTIS data base
AUTH: A/REED, W. E. CORP: National Technical Information Service, Springfield, Va.
ABS: An updated bibliography containing 140 abstracts concerning the design, performance, radiation patterns, and applications of phased arrays are presented. Applications include communications, radar, optical, spacecraft, and navigational aids.
RPT#: NTIS/PS-78/0331/5 NTIS/PS-77/0309 NTIS/PS-76/0308 NTIS/PS-75/338 78/04/00 78N29386

UTTL: Microwave oscillators. Citations from the engineering index data base
AUTH: A/REED, W. E. CORP: National Technical Information Service, Springfield, Va.
ABS: An updated bibliography containing 252 abstracts concerning bibliography cites reports from worldwide research on the design, performance, properties, and applications of microwave oscillators including Gunn, impatt, transistor, trapatt, transferred electron, Schottky, and surface acoustic wave oscillators.
RPT#: NTIS/PS-78/0334/9 NTIS/PS-77/0300 NTIS/PS-76/0313 78/04/00 78N29385

UTTL: Schlieren and shadowgraph photography. Part 2: General studies. Citations from the engineering index data base
AUTH: A/HABERCOM, G. E., JR. CORP: National Technical Information Service, Springfield, Va.
ABS: A bibliography containing 227 abstracts concerning aspects of schlieren and shadowgraph photography is presented. Techniques, equipment, and applications are reviewed for studying shock waves, combustion, acoustic waves, ballistics, and heat flow.
RPT#: NTIS/PS-78/0413/1 NTIS/PS-77/0348 NTIS/PS-76/0337 78/05/00 78N27397

UTTL: Schlieren and shadowgraph photography. Part 1: Flow visualization and measurement. Citations from the engineering index data base
AUTH: A/HABERCOM, G. E., JR. CORP: National Technical Information Service, Springfield, Va.
ABS: A bibliography containing 204 abstracts concerning

worldwide engineering research of Schlieren and shadowgraph photography techniques for flow measurement and visualization is presented. Included are studies on flames, combustion, and aerodynamic configurations.
RPT#: NTIS/PS-78/0412/3 NTIS/PS-77/0347 NTIS/PS-76/0336 78/05/00 78N27396

UTTL: Schlieren and shadowgraph photography. Citations from the NTIS data base
AUTH: A/HABERCOM, G. E., JR. CORP: National Technical Information Service, Springfield, Va.
ABS: A bibliography containing 318 abstracts on the applications and techniques of schlieren and shadowgraph photography is presented. The majority of the information is primarily concerned with flow visualization, although studies on heat transfer and combustion processes are also included.
RPT#: NTIS/PS-78/0411/5 NTIS/PS-77/0346 NTIS/PS-76/0335 NTIS/PS-75/117 78/05/00 78N27395

UTTL: Bibliographic and numeric data bases for fiber composites and matrix materials
AUTH: A/MCMURPHY, F. E.; B/QUICK, T. M. CORP: California Univ., Livermore. Lawrence Livermore Lab. Presented at 2d Conf. on Composite Material, Toronto, 16 Apr. 1978
ABS: Research leading to the creation of bibliographic and numeric data bases of material properties with particular emphasis to their application to modern flywheel technology is reported. The bibliographic data base was created to provide a direct means to visually examine pertinent literature. The numeric data base is being created to provide evaluated materials properties data for direct input to applications programs. Data bases and their evaluation programs will be stored on a PDP 11/70 computer system.
RPT#: UCRL-79503-REV-1 CONF-780414-1 78/01/11 78N27188

UTTL: Charge transfer devices, volume 3. Citations from the engineering data base
AUTH: A/REED, W. E. CORP: National Technical Information Service, Springfield, Va.
ABS: Research reports from worldwide literature are cited which cover the technology, design, fabrication, and applications of charge transfer devices. Applications include image devices, signal processors, amplifiers, filters, and memory devices. An updated bibliography containing 205 abstracts is presented.
RPT#: NTIS/PS-78/0309 NTIS/PS-77/0271 NTIS/PS-76/0241 78/03/00 78N25347

UTTL: Charge transfer devices. Citations from the NTIS data base
AUTH: A/REED, W. E. CORP: National Technical Information Service, Springfield, Va.
ABS: The technology, design, fabrication, and applications of charge transfer devices are presented in the cited Federally sponsored research reports. Applications include imaging, signal processing, detectors, filters, amplifiers, and memory devices. An updated bibliography containing 246 abstracts is presented.
RPT#: NTIS/PS-78/0307 NTIS/PS-77/0269 NTIS/PS-76/0240 NTIS/PS-75/275 78/03/00 78N25346

UTTL: Charge transfer devices, volume 2. Citations from the engineering index data base
AUTH: A/REED, W. E. CORP: National Technical Information Service, Springfield, Va.
ABS: Research reports from worldwide literature are cited which cover the technology, design, fabrication, and applications of charge transfer devices. Applications include image devices, signal processors, amplifiers, filters, and memory devices. An updated bibliography containing 239 abstracts is presented.
RPT#: NTIS/PS-78/0308 78/03/00 78N25345

UTTL: Ferrous metal casting, volume 2. Citations from the engineering index data base
AUTH: A/SMITH, M. F. CORP: National Technical Information Service, Springfield, Va.
ABS: Worldwide research is cited on iron and steel continuous casting, die casting, and precision casting. Studies on molds, mold release agents, shrinkage, solidification, defects, and additives are included. An updated bibliography containing 329 abstracts is presented.
RPT#: NTIS/PS-78/0288 NTIS/PS-77/0248 78/03/00 78N25197

UTTL: Ferrous metals casting. Citations from the NTIS data base
AUTH: A/SMITH, M. F. CORP: National Technical Information Service, Springfield, Va.
ABS: Federally sponsored research covered includes casting processes, heat treatment, quality control, solidification, microstructure, mechanical properties, additives, and purification. Studies are included on casting and its involvement in scrap reuse, air pollution, and tool manufacture. An updated bibliography containing 221 abstracts is presented.

RPT#: NTIS/PS-78/0287 NTIS/PS-77/0246 NTIS/PS-76/0142 NTIS/PS-75/223 78/03/00 78N25196

UTTL: Electrodialysis desalination. Citations from the NTIS data base
AUTH: A/CAVAGNARO, D. M. CORP: National Technical Information Service, Springfield, Va.
ABS: A bibliography containing 145 abstracts on Federally funded research presents annotated references on electrodialysis desalination, theory, membrane preparation and performance, and test and pilot plant operations. Reports are cited which cover water treatment operations as well as waste treatment processes.
RPT#: NTIS/PS-78/0243 NTIS/PS-77/0018 NTIS/PS-76/0011 NTIS/PS-75/135 78/03/00 78N25162

UTTL: Electron beam welding, volume 2. Citations from the engineering index data base
AUTH: A/REED, W. E. CORP: National Technical Information Service, Springfield, Va.
ABS: This updated bibliography contains 93 abstracts of reports from worldwide research on electron beam welding and the properties and weldability of steel, aluminum, titanium, refractory metals, and heat resistant alloys. Only three of the citations appeared in the previous edition.
RPT#: NTIS/PS-78/0186/3 NTIS/PS-77/0157 NTIS/PS-76/0137 78/03/00 78N23454

UTTL: Electron beam welding, volume 1. Citations from the engineering index data base
AUTH: A/REED, W. E. CORP: National Technical Information Service, Springfield, Va.
ABS: Reports from worldwide research are cited covering the process, automation, equipment, and applications of electron beam welding. Properties and weldability of steel, aluminum, titanium, refractory, metals, and heat resistant alloys included in this updated bibliography contain 226 abstracts, none of which are new entries to the previous edition
RPT#: NTIS/PS-78/0185/5 78/03/00 78N23453

UTTL: Electron beam welding. Citations from the NTIS data base
AUTH: A/REED, W. E. CORP: National Technical Information Service, Springfield, Va.
ABS: This updated bibliography contains 242 abstracts of federally-sponsored research reports concerning the development, automation, and applications of electron beam welding. Properties of electron beam welds and weldability of steel, aluminum, refractory metals, heat resistant alloys, and corrosion resistant alloys are included. Thirty-six citations are new entries to the previous edition.
RPT#: NTIS/PS-78/0184/8 NTIS/PS-77/0156 NTIS/PS-76/0136 NTIS/PS-75/226 78/03/00 78N23452

UTTL: Storage and retrieval of information on systems of partial differential equations and their solutions: Creatabase and the continuum/mechanics center data base of hydrocodes
AUTH: A/HIRSCHBERG, M. A.; B/LACETERA, J.; C/SCHMITT, J. A. CORP: Ballistic Research Labs., Aberdeen Proving Ground, Md.
ABS: A Continuum Mechanics Center has been established for the purposes of evaluating and developing models of interacting continua. Because of the large and growing body of literature concerning such models and related computer codes, the vast number of assumptions made in their use, and the varying types of numerical methods utilized in these codes, a data base analysis system, CREATABASE, was used to store information and characteristics of the different codes. This paper briefly describes CREATABASE, delineates the data base, describes queries made on the data base, and outlines future uses and expansion of the data base and the data base analysis system.
RPT#: AD-A050307 BRL-2015 77/09/00 78N21003

AGARD Lecture Series No.130
Advisory Group for Aerospace Research and Development, NATO
DEVELOPMENT AND USE OF NUMERICAL AND FACTUAL DATA BASES
Published October 1983
130 pages

Lecture Series No.130 is concerned with the development and use of numerical and factual data bases, and is sponsored by the Technical Information Panel of AGARD and implemented by the Consultant and Exchange Programme.

Numerical and factual data, as sources of information for all levels of aerospace and defence R & D manage-

P.T.O

AGARD-LS-130

Data bases
Information systems
Data processing
Numerical analysis
Information retrieval
Information theory
Systems engineering

AGARD Lecture Series No.130
Advisory Group for Aerospace Research and Development, NATO
DEVELOPMENT AND USE OF NUMERICAL AND FACTUAL DATA BASES
Published October 1983
130 pages

Lecture Series No.130 is concerned with the development and use of numerical and factual data bases, and is sponsored by the Technical Information Panel of AGARD and implemented by the Consultant and Exchange Programme.

Numerical and factual data, as sources of information for all levels of aerospace and defence R & D manage-

P.T.O

AGARD-LS-130

Data bases
Information systems
Data processing
Numerical analysis
Information retrieval
Information theory
Systems engineering

AGARD Lecture Series No.130
Advisory Group for Aerospace Research and Development, NATO
DEVELOPMENT AND USE OF NUMERICAL AND FACTUAL DATA BASES
Published October 1983
130 pages

Lecture Series No.130 is concerned with the development and use of numerical and factual data bases, and is sponsored by the Technical Information Panel of AGARD and implemented by the Consultant and Exchange Programme.

Numerical and factual data, as sources of information for all levels of aerospace and defence R & D manage-

P.T.O

AGARD-LS-130

Data bases
Information systems
Data processing
Numerical analysis
Information retrieval
Information theory
Systems eningeering

AGARD Lecture Series No.130
Advisory Group for Aerospace Research and Development, NATO
DEVELOPMENT AND USE OF NUMERICAL AND FACTUAL DATA BASES
Published October 1983
130 pages

Lecture Series No.130 is concerned with the development and use of numerical and factual data bases, and is sponsored by the Technical Information Panel of AGARD and implemented by the Consultant and Exchange Programme.

Numerical and factual data, as sources of information for all levels of aerospace and defence R & D manage-

P.T.O

AGARD-LS-130

Data bases
Information systems
Data processing
Numerical analysis
Information retrieval
Information theory
Systems engineering

ment and staff activity, are becoming increasingly important. These data are necessary to support research and engineering efforts in all fields. They are also becoming increasingly important to support or assist in the decision-making process. Today, a number of numerical data bases are avialable through national information centres and others are available from academic or commercial information sources. Data in many of these data bases can be retrieved and manipulated in display systems currently available. There is, however, a great need to improve the quality, reliability, availability, accesssibility, dissemination, utilization and management of these data.

Better knowledge regarding the generation and availability of such data bases, and the techniques for their use, will be of benefit to the R & D community and their information service centres.

The scope of the Lecture Series includes: generation of numerical data, consideration of the quality and reliability of the data, methods for publishing and disseminating the data, a review of the data bases that are currently available, how these data bases can be used, and future needs for numerical data bases.

ISBN 92-835-1459-9

ment and staff activity, are becoming increasingly important. These data are necessary to support research and engineering efforts in all fields. They are also becoming increasingly important to support or assist in the decision-making process. Today, a number of numerical data bases are avialable through national information centres and others are available from academic or commercial information sources. Data in many of these data bases can be retrieved and manipulated in display systems currently available. There is, however, a great need to improve the quality, reliability, availability, accesssibility, dissemination, utilization and management of these data.

Better knowledge regarding the generation and availability of such data bases, and the techniques for their use, will be of benefit to the R & D community and their information service centres.

The scope of the Lecture Series includes: generation of numerical data, consideration of the quality and reliability of the data, methods for publishing and disseminating the data, a review of the data bases that are currently available, how these data bases can be used, and future needs for numerical data bases.

ISBN 92-835-1459-9

ment and staff activity, are becoming increasingly important. These data are necessary to support research and engineering efforts in all fields. They are also becoming increasingly important to support or assist in the decision-making process. Today, a number of numerical data bases are avialable through national information centres and others are available from academic or commercial information sources. Data in many of these data bases can be retrieved and manipulated in display systems currently available. There is, however, a great need to improve the quality, reliability, availability, accesssibility, dissemination, utilization and management of these data.

Better knowledge regarding the generation and availability of such data bases, and the techniques for their use, will be of benefit to the R & D community and their information service centres.

The scope of the Lecture Series includes: generation of numerical data, consideration of the quality and reliability of the data, methods for publishing and disseminating the data, a review of the data bases that are currently available, how these data bases can be used, and future needs for numerical data bases.

ISBN 92-835-1459-9

ment and staff activity, are becoming increasingly important. These data are necessary to support research and engineering efforts in all fields. They are also becoming increasingly important to support or assist in the decision-making process. Today, a number of numerical data bases are avialable through national information centres and others are available from academic or commercial information sources. Data in many of these data bases can be retrieved and manipulated in display systems currently available. There is, however, a great need to improve the quality, reliability, availability, accesssibility, dissemination, utilization and management of these data.

Better knowledge regarding the generation and availability of such data bases, and the techniques for their use, will be of benefit to the R & D community and their information service centres.

The scope of the Lecture Series includes: generation of numerical data, consideration of the quality and reliability of the data, methods for publishing and disseminating the data, a review of the data bases that are currently available, how these data bases can be used, and future needs for numerical data bases.

ISBN 92-835-1459-9

AGARD
NATO OTAN
7 RUE ANCELLE · 92200 NEUILLY-SUR-SEINE
FRANCE
Telephone 745.08.10 · Telex 610176

DISTRIBUTION OF UNCLASSIFIED AGARD PUBLICATIONS

AGARD does NOT hold stocks of AGARD publications at the above address for general distribution. Initial distribution of AGARD publications is made to AGARD Member Nations through the following National Distribution Centres. Further copies are sometimes available from these Centres, but if not may be purchased in Microfiche or Photocopy form from the Purchase Agencies listed below.

NATIONAL DISTRIBUTION CENTRES

BELGIUM
Coordonnateur AGARD – VSL
Etat-Major de la Force Aérienne
Quartier Reine Elisabeth
Rue d'Ev[illegible] 1140 Bruxelles

CANADA
Defenc[illegible]
Depart[illegible]
Ottaw[illegible]

DENMARK
Dani[illegible]
Øste[illegible]
Cop[illegible]

FRANCE
O[illegible]
29[illegible]
9[illegible]

GERM[illegible]

GR[illegible]

ICELAND
Director of Aviation
c/o Flugrad
Reykjavik

ITALY
Aeronautica Militare
Ufficio del Delegato Nazionale all'AGARD
3, Piazzale Adenauer
Roma/EUR

LUXEMBOURG
See Belgium

NETHERLANDS
Netherlands Delegation to AGARD
[illegible] Aerospace Laboratory, NLR

[illegible]ishment

[illegible]ate

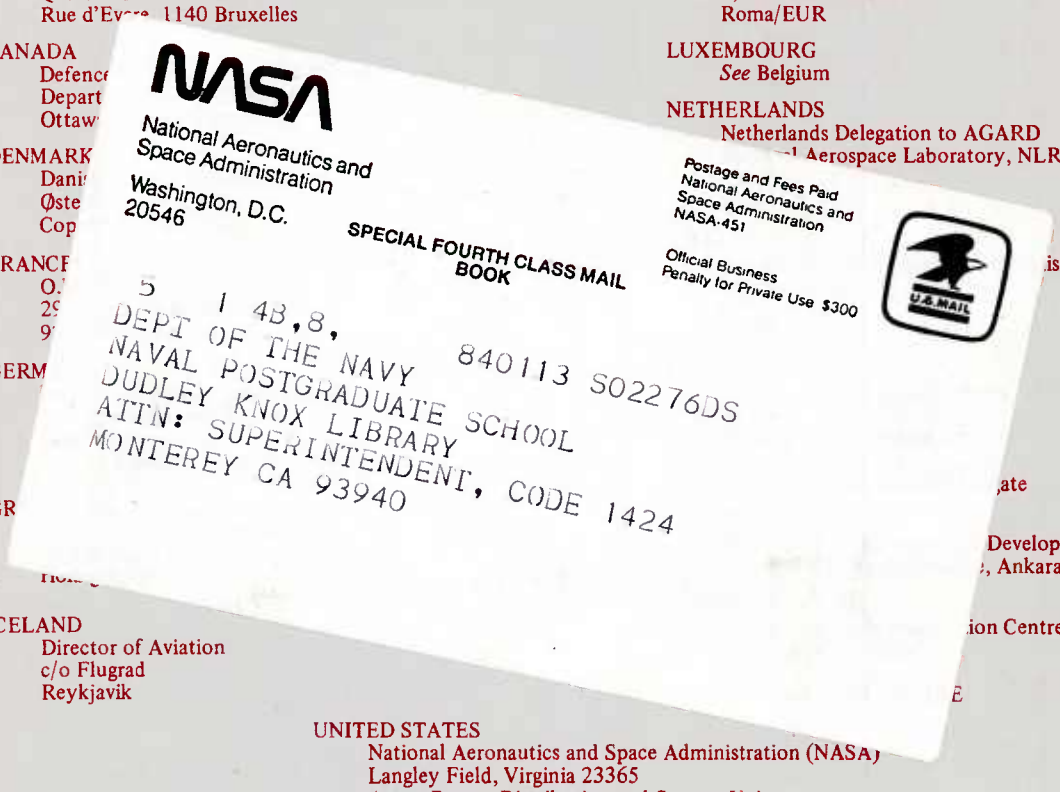
[illegible] Development (ARGE)
[illegible], Ankara

[illegible]ion Centre

[illegible]E

UNITED STATES
National Aeronautics and Space Administration (NASA)
Langley Field, Virginia 23365
Attn: Report Distribution and Storage Unit

THE UNITED STATES NATIONAL DISTRIBUTION CENTRE (NASA) DOES NOT HOLD STOCKS OF AGARD PUBLICATIONS, AND APPLICATIONS FOR COPIES SHOULD BE MADE DIRECT TO THE NATIONAL TECHNICAL INFORMATION SERVICE (NTIS) AT THE ADDRESS BELOW.



PURCHASE AGENCIES

Microfiche or Photocopy
National Technical
Information Service (NTIS)
5285 Port Royal Road
Springfield
Virginia 22161, USA

Microfiche
ESA/Information Retrieval Service
European Space Agency
10, rue Mario Nikis
75015 Paris, France

Microfiche or Photocopy
British Library Lending
Division
Boston Spa, Wetherby
West Yorkshire LS23 7BQ
England

Requests for microfiche or photocopies of AGARD documents should include the AGARD serial number, title, author or editor, and publication date. Requests to NTIS should include the NASA accession report number. Full bibliographical references and abstracts of AGARD publications are given in the following journals:

Scientific and Technical Aerospace Reports (STAR)
published by NASA Scientific and Technical
Information Branch
NASA Headquarters (NIT-40)
Washington D.C. 20546, USA

Government Reports Announcements (GRA)
published by the National Technical
Information Services, Springfield
Virginia 22161, USA

Printed by Specialised Printing Services Limited
40 Chigwell Lane, Loughton, Essex IG10 3TZ

ISBN 92-835-1459-9